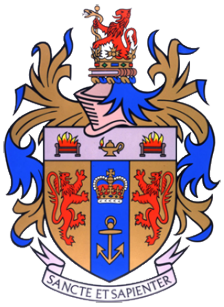
Structure Based Online Social Network Link Prediction Study

Gao, Fei

Awarding institution:
King's College London

Structure Based Online Social Network Link Prediction Study



Fei Gao

Department of Informatics

King's College London

A thesis submitted for the degree of

Doctor of Philosophy

1 November 2017

Acknowledgements

It would not have been possible to complete this thesis without the support of a number of people.

First and foremost, I want to express my appreciation to my supervisor Dr Katarzyna Musial-Gabrys, who kindly guided me throughout my Ph.D. research and taught me the art of social network analysis. Her enthusiasm, encouragement, and faith in me have been extremely helpful, especially during the tough time. It has been great to work with her because she always has plenty of inspiring ideas.

Meanwhile, I would also like to thank Professor Peter McBurney, Professor Colin Cooper and Dr Sophia Tsoka for their help and support in the past four years, especially during the second year when Dr Katarzyna Musial-Gabrys was away for maternity leave. I appreciate it very much.

In regards to the Ph.D. life, I would like extend my thanks to all my colleagues at KCL, Emre Savas, Santhilata Kuppili Venkata, Jonathan Silva, Jia Gao, Wiktor Piotrowski, as well as those who have already joined the Dr club, Dr Shuyu Ping, Dr Yansha Deng, Dr Junhuan Zhang, the best couple Dr Samhar Mahmoud and Dr Lina Barakat, Dr Michal Sroka, Dr Bram Ridder and Dr Michael Cashmore. All of you make my life as a Ph.D. student colourful.

I would never forget the fun time we had playing board together during lunch.

I would also like to take this chance to thank my manager at FactSet, Ramon Gonzalez. Thank you for your support and approving my holiday and work from home request during my thesis writing period.

Last but not least, I would like to thank my parents for all their love and unconditional support. I also acknowledge my wife, Yan, who is my champion and blessed me with a life of joy in the hours when the office lights were off.

I love you all.

Abstract

This thesis shed light on the Internet-based social network link prediction problem. After reviewing recent research achievements in this area, two hypotheses are introduced: (i) The performance of topology- based network prediction methods and the characteristics of the networks are correlated. (ii) As networks are dynamic, the performance of prediction can be improved by providing different treatment to different nodes and links. To verify the Hypothesis (i), we conduct experiments with six selected online social networks. The correlation coefficients are calculated between six common network metrics and ten widely used topology-based network link prediction methods. The results show a strong correlation between Gini Coefficient and Preferential Attachment method. This study also reveals two types of networks: prediction-friendly network, for which most of the selected prediction methods perform well with an AUC result above 0.8, and prediction unfriendly network that on the contrary. For Hypothesis (ii), we proposed two network prediction models, the Hybrid Prediction Model and Community Bridge Boosting Prediction Model (CBBPM). The hybrid prediction model assumes network links are formed following different rules. The model linearly combines eight link prediction methods and the evolvement rules have been probed by finding the best weight for each of the method by solving the linear optimization problem. This experiment result shows an improvement

of prediction accuracy. This model takes link prediction as a time series problem. Different from Hybrid Prediction Model, CBBPM provides a different treatment on nodes. We define and classify network nodes as community bridge node in a novel approached based on their degree and links position in network communities. The similarity score that calculated from the selected prediction methods is then boosted for predicting new links. The results from this model also show an enhancement of prediction accuracy. The two hypotheses are validated using the research experiments.

Abbreviations

CC - Clustering Coefficient

GCC - Global Clustering Coefficient

LCC - Local Clustering Coefficient

ASP - Average Shortest Path

CN - Common Neighbours

JI - Jaccard's Coefficient Index

PA - Preferential Attachment

AA - Adamic/Adar Index

RA - Resource Allocation

Cos - Cosine Similarity

Sor - Sørensen Index

Katz - Katz$_\beta$ Method

HPI - Hub Promoted Index

HDI - Hub Depressed Index

LHNI - Leicht-Holme-Newman Index

AUC - Area Under the Receiver-Operating Characteristic

CBBPM - Community Bridge Boosting Prediction Model

BNSS - Bridge Nodes Similarity Score

MCDR - Max Community Dominant Rate

Contents

List of Figures

Chapter 1

Introduction

This chapter described the background of the work. The search problem is formalised and then the research questions and objectives are introduced. The whole thesis structure is introduced in the end of this chapter.

1.1 Background

Networks have been studied for a long time and their origins can be traced back to 1736 when Euler defined and solved the Seven Bridges problem of Königsberg [1]. Since then, for a long time, networks have mainly been studied by mathematicians within the scope of graph theory. There was no significant progress in complex network research until 1960s, when the Erdos-Renyi random graph model (ER-model) was introduced [2; 3]. This is the simplest model of complex networks. However, due to the fact that there was lack of large-scale real world data that could be used for the network research, more efforts was made in the direction of theoretical analysis. The richness of network information, such as the links and nodes attribute information, is also not comparable to what we have today as data collection and recording technology has improved a lot. During the time when ER-model was introduced,

progress has also been made by sociologists in researching real world human relationships, e.g.[4; 5]

A new wave of research was set off by Watts and Strogatz who published a paper about the small-world effect in Nature in 1998 [6] and introduction of the scale-free network models by Barabasi and Albert one year later [7]. With the expansion of the Internet, more and more real world network datasets are available for research and the network information is much richer and more complex than before. Complex network, which can be studied as an abstract form of various networks, has attracted the attention of scientists who focus on the real networks including study of biological networks such as protein-protein interaction networks [8; 9; 10; 11], metabolic networks [12; 13; 14; 15], work on epidemic disease spread among human networks [16; 17; 18; 19], research on scientific collaboration networks [20; 21; 22; 23], neural networks [24] or study of online social networks [25; 26; 27; 28].

Social networks have also been studied for many years. However, rapid development of the Internet in the past few decades has pushed the research in the area of network science to an entirely new level. More and more human activities have been moved from off-line to on-line world and this resulted in vast amount of data available for investigation. Online social networks, ranging from collaboration networks to friendship networks, and from networks obtained from phone calls to email communication networks, have been widely studied by researchers from various areas. Generally, these social networks can be represented as graphs where nodes are users and links indicate social interactions between those users. Triggered by human activities, social network keeps changing which makes the network prediction a challenging and worth studying topic.

Prediction of complex network is one of the popular research topics in the realm of network science. Most of the researchers focus on the link prediction

problem [29; 30] which is very valuable for solving real world problems. Generally, the link prediction problem is mainly studied from two perspectives: (i) topological information and (ii) information about attributes of nodes and edges.

Topology refers to the arrangement of nodes and links that compose the network. It reflects the structural information of networks. In the context of social network, nodes are the users and links refer to the relationship between them. It could be friendship, or communication connections like message and email. The link prediction is to predict the new links between nodes based on their existing network topology information. The classic approach for solving the link prediction problems is first to take a snapshot of a network resulted from the time frame $[t_0, t_1]$. New links are predicted based on the network topology information available in $[t_0, t_1]$. The results are verified with the real world network snapshot from the time range of $(t_1, t_2]$. Algorithms for links prediction typically compute similarity score between two nodes and assume that nodes with larger scores are more likely to be connected. Most of the achievements made on structure based prediction are by mathematicians and physicists. Some of the well-known structure based prediction methods[31] are Common Neighbour, Jaccard's Index, Adamic/Adar Index, Katz.

The prediction problem also been studied from the angles of the network attribute information. The attribute information refers to extra descriptions about the detail features of each node. Such information is difficult to show directly in the network graph, it is usually presented in a form of database tables. Majority of attribute based prediction methods follow a machine learning approach, use classifier to do predict based on part of the whole dataset (which usually called training data) and using the rest data to verify the prediction result. Widely used methods include Decision Tree, Support Vector Machine (SVM), Naïve Bayes [32; 33].

However, although lots of efforts have been made, there is still a huge room for research.

1.2 Motivation

This work focuses on the link prediction problem which has been formalized in [29]. The motivation of this research can be summarized in three main points.

First, the study of complex network prediction can help researchers to gain a better understanding of the evolution of complex networks. Many efforts have been made to solely study the dynamics of complex network [34; 35; 36]. However, the achievements of network prediction research, can help to explain the mechanisms that drive network evolution.

Second, [29] show that the performance of structure based prediction methods vary greatly. This trend also exists in attribute based prediction methods. [37] studies the scientific co-author network with an outcome that SVM performance is the best among the methods used in the experiment. Another research by [32] focuses on mobile social networks and in this case it turned out that the Decision Tree and Logistic Regression are the best performing approaches in the conducted experiments. The performance of the prediction methods are highly network depended. Previous researches always focus on the performance of methods on one or few networks without considering the correlation between the methods and network characteristics. There is a need to analyse the this relationship in order to design a prediction framework that will enable to achieve better performance and this also motivated presented in this thesis research.

Third, the research on prediction of complex network is also important for many other subjects. The prediction could help finding the potential protein relations which might not be easily observed directly due to the complexity of protein-protein interaction network. For example, new interactions can be inferred from the existing known interaction networks [38; 39] which shows a much better performance than prediction purely by chance. On-line market

targeting might also benefit from the complex network prediction which has already been applied in real world industries. For example, Google and Amazon recommends customers the potential goods that they might be interested which is a kind of link prediction that predict the link between customers and products. Friends can also be predicted for users in social network [40]. What's more, the complex network theory could also benefict traffic control area [41].

1.3 Problem Statement

Social network, as a type of complex network, provides wealth of network information for study. Previous work on social network prediction has mainly focused on general information of network such as degree and distance of nodes which can be collected directly from the network snapshot structure [29]. The approach used by mathematicians and physicists assume that all the nodes are the same and can be seen as particles[31]. Structure analysis reveals several very important social network properties such as short path length, high clustering and power-low degree distribution[42]. However, the real world social network keeps evolving which makes each node and link different from others in the network. **A dynamic prediction approach that treats nodes or links differently could be a good try to help improving the prediction performance.**

1.4 Research Questions and Objectives

The research goal is to propose a new prediction method for large social networks which could provide a better prediction accuracy than existing approaches. To achieve the goal, we need to have a systematic analysis of the existing prediction methods.

One of the main objectives of my PhD research is to understand the relationship between the network link prediction methods and complex social network characteristics. Here are two hypothesises that we made in this research:

• **Hypothesis 1**: The performance of topology based network prediction methods and the characteristics of the networks are correlated.

For the topology based methods (such as common neighbour or Katz), all the information needed for the prediction is extracted from the network structure. The prediction accuracy varies a lot depending on the network to which a given method is applied [31]. One possible reason could be that the accuracy of topology based methods is related to the characteristics of network structures.

• **Hypothesis 2**: As network are dynamic, the performance of prediction can be improved by providing different treatment to nodes and links.

It has shown in [29] that topology based methods are much better than the random prediction. All the topology based prediction methods are based on their own assumptions. For example, the common neighbour method assumed the more common neighbour the two nodes have, the more likely the two nodes

would form new link. If a network keeps changing based on this assumption, then common neighbour should outperform all the other network prediction methods and that is the rule of how the network evolves. This focuses more on the nodes treatment in network. Meanwhile, the network structure information, such as network community, are not considered in the topology based methods while they also play an important role in reflecting network evolvement. For example, the links within and inter the communities are not same in forming the latest network structure. This should also be considered when predicting new links.

These two hypothesises are the expected outcome of this research and also based on them, we can summarize our research questions:

1. What is the relationship between the performances of topology based prediction methods and the characteristics of networks (such as degree distribution, clustering coefficient etc.)?

2. What treatment should be given to different nodes and links based on the characteristics that the network's components have?

The research goal can be achieved by finding the answer to the questions and meanwhile, the hypothesises could also be verified. To achieve this, we established the following objectives:

1. To obtain real-world social networks and predict links using widely used topology based methods.

2. To investigate the correlation between the link prediction accuracy and network metrics;

3. To propose a hybrid prediction approach which could provide different treatment for networks evolving in different ways.

4. To develop a method that could predict new links with considering the network structure information.

1.5 Thesis Structure

This thesis is structured as follows: Chapter 2 reviews the basic network concepts, network model as well as the structure based link prediction methods. Then the metholody for my research is introduced in Chapter 3. Following that, Chapter 4 introduce all the networks we used for the experiments performed in this thesis. This chapter also described in detail about how each selected network is processed for different study. To verify the Hypothesis 1 stated in Chapter 1.4, Chapter 5 introduces the correlation study between link prediction accuracy and network metrics. The two models, hybrid model and community bridge boosting prediction model, are described and tested in Chapter 6 and Chapter 7. Finally, all the works are summarized and potential future works are stated in the end Chapter 8.

Chapter 2

Literature Review

This chapter reviews existing studies and achievements that related to the link prediction problem. Useful background knowledge, such as basic network concepts and network models, are also reviewed in this chapter.

2.1 Basic networks concepts

In the recent years, complex networks have become more and more popular research topic. In the context of network theory, a complex network is a graph (network) with significant topological features that do not occur in simple networks such as lattices or random graphs but often occur in real graphs [42]. For a complex network, the global characteristics cannot be directly inferred from local information [42; 43]. Although each individual in the network follows simple rules in the network, the behaviour of the network as a whole could be very different which is also known as emergence phenomenon [44]. The research of complex networks has attracted researchers from various communities which include social science [19; 25; 45; 46], physics [47; 48; 49; 50], biological science [12; 13; 51; 52] and computer science [29; 31; 53; 54].

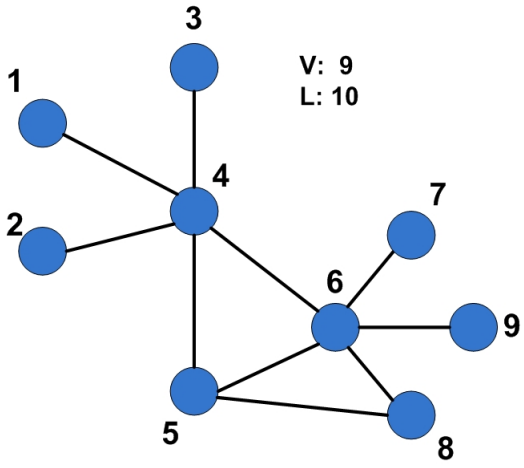


Figure 2.1: Network Example

Each network consists of sets of nodes and links between nodes. Figure 2.1 shows a very basic network with 9 nodes and 10 links. In real world, the node can be a person in social network, a router in internet network, a web page in WWW, etc. The link can represent a relationship between two persons, a connection between routers, a link referring from one web page to another, etc. For example, in [21], nodes in a network are authors who contribute to a Wikipedia page. The link between nodes imples that two authors(nodes) have worked together on the same page. In [12], each node represents a kind of protein and the link between them means direct physical interactions.

The study of networks has a lot of practical meanings, especially to the current world where there are many problems. For instance, terrorism has become a worldwide issue nowadays. Terrorists use network to communicate and broadcast radicalism to general public [55]. To prevent them from bringing more disasters to human society, there is an urgent demand to gain understanding about how information spreads in network as well as how to predict violent behaviour. Thus, network analysis is a very important topic.

Network analysis is useful for people from different fields. For the terrorism

issue as mentioned in last paragraph, network analysis could help governments with law enforcement [56]. It could also help police and national defence department to target criminal networks [57]. Researchers also apply network analysis in cancer study which provides them a roadmap to investigate new diagnostic and therapeutic opportunities across cancer types [58]. Thus, network analysis could benefit lots of researchers.

Types of Networks

The network in Fig 2.1 is an example of binary unweighted network. The links between the nodes have no directions and all links are same without weight. i.e. they either exist or not. When considering these two features: (i) direction and (ii) weight, networks can be classified into four main categories. Undirected and unweighted network is shown in Figure 2.2(a). Relationships in this tye of network are undirected and binary with only two conditions exist or not. Directed and unweighted network is shown in Figure 2.2(b) where relationships between nodes are binary and have direction. The direction means, for example, one user follows another user in twitter network. Another example is undirected and weighted network as shown in Figure 2.2(c). The link in this type of network does not have direction but has weight as the number shown beside each link. The weight can reflect strength of the links. Directed and weighted network is shown in Figure 2.2(d), which can be seen as the combination of previous two networks, where relationships have both direction and weight. Last type of network is the most complex case.

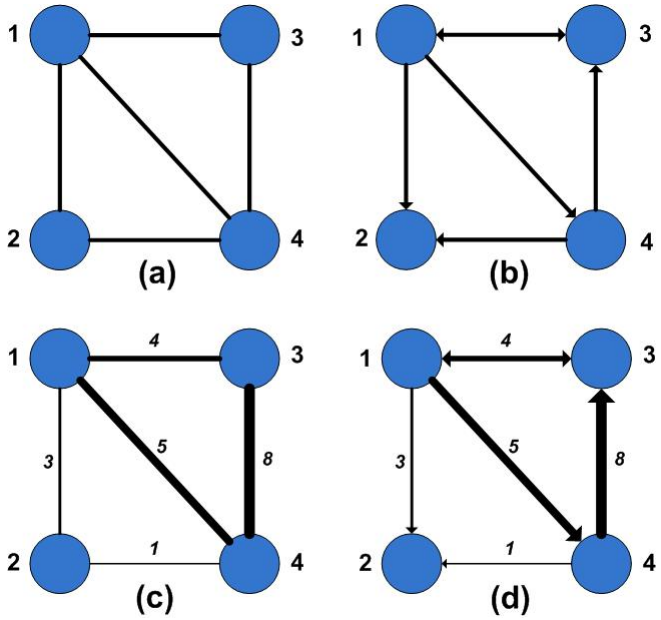


Figure 2.2: Four Types of Networks

2.1.1 Basic Network Metrics

Below, basic network metrics, used to describe the networks, are introduced. This include node degree, degree distribution, common neighbour concept, node distance, average shortest path, and clustering coefficient.

Node Degree

Degree of a node is the number of edges incident to a given node (loops are counted twice) [59]. For example, in Fig 2.1, the degree of node 6 is 5, the degree of node 8 is 2. In this thesis, k is used to describr degree of a specific node i.

Node Degree Distribution

Degree distribution is the probability distribution of all the nodes degrees over the whole network [59]. Thus, it reflects the global property of the network. It is a very important factor to when studying both real-world and theoretical

networks as it provides a lot of information about networks type and main networks characteristics. For example, if the degree distribution of a network follows power law, it is a scale-free network [Section 2.3.4] that has some hub nodes meaning that the structure is not resilient for targeted attacks.

Common Neighbour

A neighbourhood of a node is a set of nodes that directly connect to this node by link [59]. By default, neighbour means first order neighbour. There is also second order neighbour which refers to the neighbours of the neighbours of a given node. Similarly, one could also define third or fourth order neighbours. Common neighbours are the neighbours that are shared by a pair of nodes. In Fig. 2.1, the common neighbours of node 4 and 8 are 5 and 6.

Node Distance

The distance of two nodes is the number of links in shortest path connecting them[59]. For example, in Fig. 2.1 the distance of node 4 and node 7 is 2. The distance that is most commonly used is average shortest path (ASP) 2.1.

$$ASP = \frac{1}{n \cdot (n-1)} \cdot \sum_{i!=j} d(i,j)(i,j \subset n) \tag{2.1}$$

Where n is the number of nodes in the network. i and j are nodes from network nodes set and $d(i,j)$ stands for the shortest path between i and j.

Clustering Coefficient

Clustering Coefficient (CC) is a concept in graph theory and it is used to measure the degree to which the nodes in a network tend to cluster together [59]. There are two types of Clustering Coefficient.

Local Clustering Coefficient(LCC) defines the CC index of one specific

node. It was defined by Duncan J. Watts and Steven Strogatz in their research on small world networks as introduced in Section 2.3.3 [6]. Figure 2.3 shows the LCC example for sample subgraphgs in a network. LCC is calculated as:

$$LCC_i = \frac{NLN_i}{TNLN_i} \tag{2.2}$$

Where NLN_i is the number of links exists between the neighbours of the node i. $TNLN_i$ is the total number of links that could exist between the neighbours of node i.

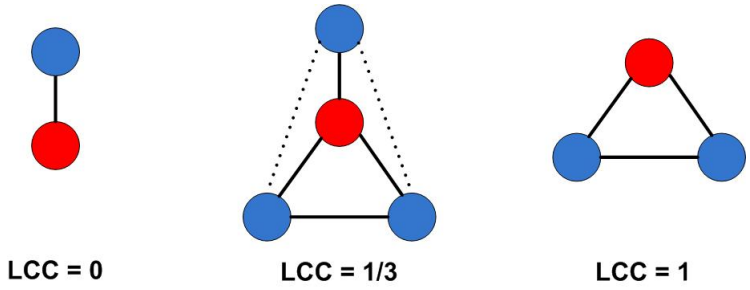


Figure 2.3: LCC Example

Global Clustering Coefficient (GCC) shows the cluster degree for a whole network. It is the number of closed triplets (or 3 x triangles) over the total number of triplets (both open and closed) which given by this:

$$GCC = \frac{3 \cdot \text{Number of Triangles}}{\text{Number of connected triplets of Nodes}} \tag{2.3}$$

In this fomular, number of triangles refer to a closed triangles with three nodes and three links. The connected triplets of nodes are the structures formed by three nodes with two links connecting them.

2.2 Social Networks

A social network is a group of people with some forms of contacts or interactions between them[60]. Traditional social networks that have been studied includes acquaintance network[5; 61], sexual contacts networks [62; 63; 64], criminal community network [57; 65; 66] and mobile social networks [67; 68; 69]. The research has revealed several real world social network characteristics [5; 6; 70]. Social networks are highly clustered meaning that they have high LCC and GCC. The network diameter is small with an average of 6 degrees which reflects small world phenomenon (section 2.3.3) and majority of social networks follow a power law degree distribution (section 2.3.4).

2.3 Network Models

Lots of studies have been done in this area in recent years as there are more and more available complex networks that could be used for study. The empirical study of complex network has shown that real word networks have some significant properties such as small-world effect and power-law degree distribution. To get explanations for observed network properties, various complex network models are introduced. The following section will introduce some classic network models.

2.3.1 Regular Network Model

Regular network is a type of network that is highly ordered. Lattice network (Figure 2.4) and circular lattice (Figure 2.5) are two typical examples of regular network models. All nodes in a regular network have exactly the same number of links (the same degrees). For a regular network, there is no need

to do statistical analysis as the nodes are distributed uniformly with the same patterns of connections.

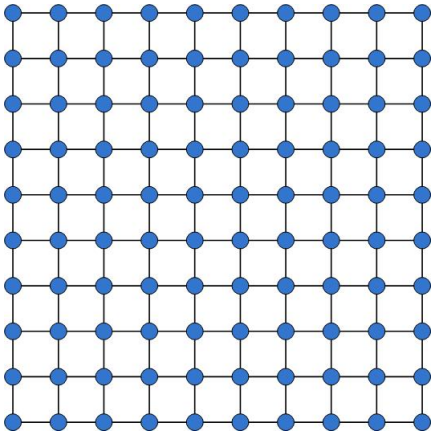


Figure 2.4: Lattice Network

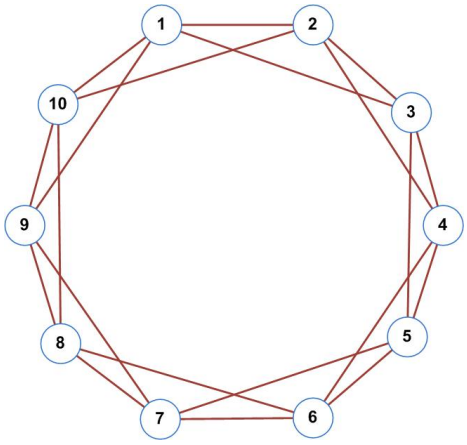


Figure 2.5: Circle Lattice

As shown in Figure 2.4 and Figure 2.5, they are all ordered networks without any trace of randomness. Those networks have a large diameter and offer low connectivity meaning that the shortest path length between two randomly chosen nodes is very long. The consequence of such characteristics is that regular networks ar not a good representation of social networks.

Advantages and Disadvantages of Regular Network Models

Advantages

◇ Regular network is usually an artificial network, thus it can be generated according to specific needs of researchers.

◇ Regular network has the lowest heterogeneity. The number of connections each node has is more or less the same. Meanwhile, the randomness of the network is also the lowest. That means the probability of any two randomly chosen nodes to be linked directly to each other is very low. This helps researchers undertand some extreme network characteristics given certain number of nodes and links such as average shortest path and clustering coefficient.

Disadvantages

◇ In real world, it is rare to observe any regular network. Thus, the benefit of studing regular network is limited for understanding human social networks.

2.3.2 Random Network

Generally speaking, random network is a network model that build a network randomly with some specific parameters taking fixed values . For instance, given a fixed number of nodes n and fixed number of links e, many networks with different connections can be created [59]. A random network example can be find in 2.6 as the right most graph.

Erdős-Rényi Model

Erdős-Rényi Model (ER) is the most fundamental and widely studied network model which is also known as Poisson random graph or Bernoulli random graph because its node degree distribution follows the Poisson distribution [3]. It is a model with fixed number of nodes n and a fixed probability p of the edges between nodes. This model is given as $G(n, p)$ where n and p stands respectively for the number of nodes and the probability that an edge exists between two randomly selected nodes. Some basic properties of ER model [2; 3; 59] are showed in Table 2.1. The A in Diameter is a constant that independent of n.

Expected Total Number of Edges	**Mean Degree**	**Clustering Coefficient**	**Diameter**
$e_{exp} = \binom{n}{2} p$	$\langle k \rangle = (n-1)p$	$C = \frac{\langle k \rangle}{n-1}$	$D = A + \frac{\ln n}{\ln \langle k \rangle}$

Table 2.1: Properties of the ER model.

Giant Component

The giant component is the component in a network that size grows in proportion to the total number of nodes n [59]. For a Poisson random graph, the condition that a giant component exists is $np > 1$. At $np \leq 1$, there is no giant component.

Small Component

The remainder of the network except giant component is made up of many small components whose average size does not increase with increase of the total number of vertices of the network [59].

Configuration Model

Configuration model is a random graph with a given degree sequence rather than degree distribution. It is given as $G(n, m)$ where n is the total number of the network nodes and m is the total number of edges. One can think that this model gives a fixed number of nodes with fixed number of stubs on each of them at beginning. To form network, these nodes need to connect to other nodes via the stubs to form full links. The random part of this model is how those stubs are connected as there could be more than one way for them to connect and form a network. Some other properties of configuration model are showed in Table 2.2. The number of stubs is $2m$ as each link can be break in

the middle to form two stabs. In the edge probability equation, k_i and k_j are the degree of nodes i and j.

Number of Stubs	**Edge Emerging Probability**
$2m$	$\frac{k_i k_j}{2m-1}$

Table 2.2: Properties of the Configuration Model

Excess Degree Distribution

The excess degree of the node is the number of edges attached to a node other than the edge we arrived along. It is equal to the total degree of the node minus one. The probability that we reach a node of degree k upon following an edge in this way is proportional not to p_k but to kp_k [70]. This is a property specific to configuration model. In real world, the degrees of adjacent vertices are often correlated. The probability of reaching a node of degree k when we follow an edge depends on what node we are coming from. Configuration model reflects this property which is a reason why this model is useful for understanding the world around us.

Advantages and Disadvantages of the Random Network Models

Advantages

◇ Random graph models have been studied in the field of network analysis for many years. This means their characteristics and processes are very well understood.

◇ Random network model reflects the small world effect. The small-world effect is the observation that the geodesic or shortest-path distance between most pairs of venodesrtices in a network is small typically just few steps away

even in networks with billions of nodes such as the acquaintance network of the entire world population[59].

◇ It is very flexible so can be defined by researcher for different purpose, e.g researchers can adjust the connection probability according to the real world situation.

Disadvantages

◇ It shows essentially no transitivity or clustering. For example, the real world social network have communities and group properties or in another words high clustering which is not found in most of the random network models.

◇ The shape of random graph degree distribution is different from the real world networks.

2.3.3 Small-world Network Model

The random network (Poisson random network and configuration model) cannot reflect the high transitivity of real network and the circle lattice cannot reflect the short path length. Each of those two models reflects one aspect of real world networks: (i) random network exhibits short average path length and (ii) regular network high clustering coefficient. To combine these two properties in one model, small-world model was proposed.

Watts and Strogatz Model

The Watts-Strogatz model, as shown in Figure 2.6, is a random graph generation model that produces graphs with small-world properties which means it

has a small average shortest path length and large clustering coefficient. The process of obtaining small-world network begins with a regular ring lattice. Then the edges are rewired with a fixed probability p. Table 2.3 states basic properties of Watts and Strogatz model include clustring coefficient and average path length

Clustering Coefficient	**Average Path Length**
$C(p) = \frac{3(k-2)}{4(k-1)}(p=0)$ $\Rightarrow \frac{k}{n}(p=1)$	$l(p) = \frac{n}{2k}(p=0)$ $\Rightarrow \frac{\ln n}{\ln k}(p=1)$

Table 2.3: Properties of the Watts and Strogatz Model

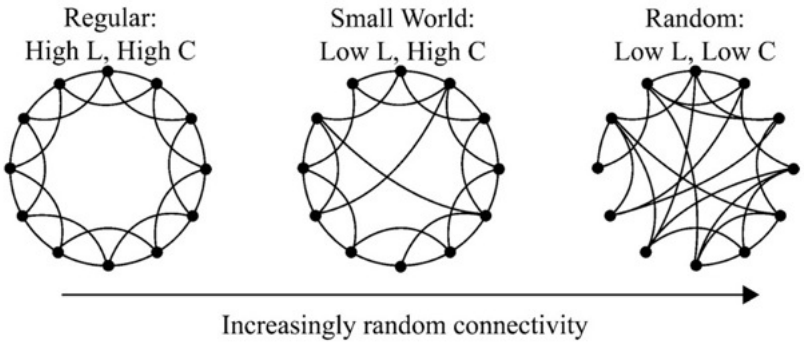


Figure 2.6: Watts and Strogatz Model [6]

Newman and Watts Model

This is a variation of the Watts and Strogatz Model model. In this model, instead of rewiring links, new links, which are also called shortcut links, are added randomly. Thus, no links are removed from the graph. This model is somewhat easier to analyse than the original one as it is not possible for any region of graph to become disconnected from the rest.

Advantages and Disadvantages of the Small-world Network Model

Advantages

◇ Small world network is used in many areas such as sociology, earth sciences and computing because it reflects the small world phenomenon which are widely existing in real world networks.

◇ It captures both the properties of high transitivity and short path length.

Disadvantages

◇ The Watts-Strogatz model and Newman-Watts modle imply a fixed number of nodes thus cannot be used to model network growth.

2.3.4 Scale-free Network

Scale-free networks are defined by power-law node degree distribution. Lots of real-world networks follow the distribution in equation 2.4 where $\alpha is an constant$.

$$P_k = \frac{1}{k^\alpha}(\alpha > 1) \tag{2.4}$$

Preferential Attachment Model (Price's Model)

The essential idea of Preferential Attachment model (PA model) is 'rich-get-richer' [59]. Price's model is as follows. We can assume that papers are published continually and the newly appearing papers cite previously existing ones. As no paper ever disappears after it is published, the nodes in this network are created but never destroyed and it is a directed edge network. The crucial central assumption of Price's model is that a newly appearing paper cites previous ones chosen at random with probability proportional to the number of citations those previous papers already have which reflect the idea of "rich-get-richer".

This model is useful because it generates a power-law degree distribution which is similar to that observed in real networks. Table 2.4 shows basic properties of Price's model. In the equation, k is the number of citation of the paper in the model, every paper are given an initial citation of 1. Because the model is very specific and limited to the citation process, which could be quite different from the growth of other networks, it cannot be generalised. Meanwhile, as for a citation process, this model also omits many other factors in the real world citation such as the change of citation number of each paper as well as the shift of research hot tops in academia. It could only reflect the fundamental mechanism behind the observed power-law degree distribution.

Mean Degree	**Degree Distribution**
$\sum_k kP(k) = \langle k \rangle$	$P(k) = \frac{B(k+1, 2+\frac{2}{m})}{B(1, 1+\frac{2}{m})}$ $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ $\Gamma(\alpha) = \int_0^\infty x^{a-1}e^{-x}dx$

Table 2.4: Properties of the PA Model.

Barabási-Albert model

It is a model for generating random scale-free networks using a preferential attachment mechanism. In this model, vertices are added one by one to a growing network and each node connects to a suitably chosen set of previously existing vertices. The node with larger degree would be given a higher probability that new coming nodes will be connect to it. The connections, however, are undirected and the number of connections made by each node is a constant number denote as c.

The Barabási-Albert model (BA model) is one of the well-known generative network models. Different from Price's model, this model is an undirected and the number of connections made by each node is an exact number c while in

Price's model, only the average value of c is fixed. The BA model generates a degree distribution with a power-law tail that always has an exponent $\alpha \approx 3$. Some other properties are showed in Table 2.5. The clustering coefficient and mean degree is relevant to the number of nodes while the degree disctibution is relevant to average node degree k. In the formular to calculating probability of new node connected to node i, k_i is the degree of node i and the sum is made over all pre-existing nodes j [71].

Clustering Coefficient	**Mean Degree**	**Degree Distribution**	**Probability of New Node Connected to Node i**
$CC \sim n^{-0.75}$	$\frac{\ln n}{\ln(\ln n)}$	$P(k) \sim k^{-3}$	$p_i = \frac{k_i}{\sum_j k_j}$

Table 2.5: Properties of the BA Model

Node Copying Models

Node copying models use copying mechanism to simulate the formation and growth of a network. In the general copying model, a growing network starts as a small initial graph and, at each time step, a new node is added with a given number c of new outgoing edges. As a result of a stochastic selection, the neighbours of the new node are either chosen randomly among the existing nodes or one existing node is randomly selected and w of its neighbours are 'copied' as heads of the new edges.

Advantages and Disadvantages of the Scale-free Network Model

Advantages

◇ The degree distribution follows a power law which can be found in many real world networks.

◇ There is a lot of real world data can be used for scale-free network research

and verification.

◇ It also reflects the small world phenomenon.

Disadvantages

◇ Although many real-world networks are thought to be scale-free, the evidence often remains inconclusive. Thus, the scale-free nature of many networks is still being debated by the scientific community.

2.3.5 Discussion

Random network model constructs network with the assumption that links between nodes are set up randomly. Although this model shows the small world effect which is also observed in real world social networks, the randomly growth of links make the prediction performance of any other link prediction methods no better than the random link prediction. This situation also happens in small world network model where links are rewired or added randomly. The purpose of the model is to simulate a process of forming a high clustered network with low diameter. It cares more about reaching the status of small-world network rather than the growth of the network. In this respect, small-world model is of limited applicability in link prediction problem. Scale-free network model is a growing network model that could generate a network with power law degree distribution. The growing mechanism is 'rich get richer'. New links are formed only when new nodes are added. However, in link prediction problem, links are also predicted between existing nodes. Scale-free network model is not suitable for classical link prediction study where only links not nodes are added, but it might be useful in node prediction where focus is on predicting links between existing nodes and new coming nodes.

To sum up, those three network models help researchers gain a better un-

derstanding about the network formation and characteristics. However, for the network prediction problem, those models can only provide background knowledge about networks and also criteria for network classification. They are only first step in the whole process of understanding network dynamic and prediction.

2.4 Structure-based Network Prediction

The networks shown in Figure 2.2 are static. However, in the real world, the shape and size of networks usually keep changing. The prediction of network evolution has a significant practical meaning. For instance, the prediction of the network evolution could help build a more efficient online recommender system (e.g. friend recommendation in Facebook and items recommendation in Amazon) or analysis the disease spread around human society (e.g. Predict and prevent the spreading of infectious diseases in socially structured populations [72]).

Link prediction problem has been widely studied in complex network community. David Liben-Nowell and Jon Kleinberg has formalized the link prediction problem in [29]. Researchers from physics and maths communities approach the problem by focusing on the topology information about the network. There are three types of link prediction problems that will be formalized in this section.

Adding Links

Adding links focuses on the link prediction that a new link will be created between existing nodes in the next time window (Figure 2.7). There can be one or more new appearing links.

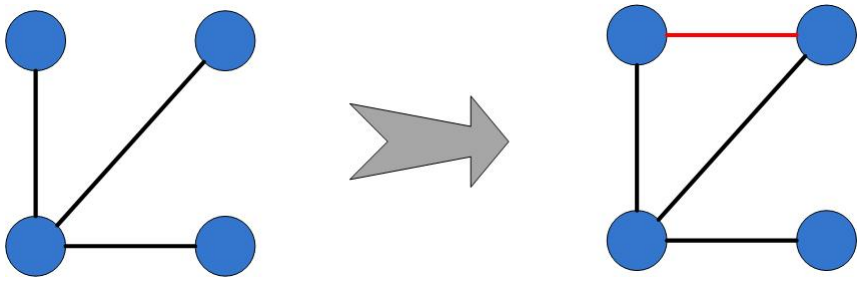


Figure 2.7: Adding Link Prediction

Removing Links

Removing links problem focuses on the predictions that a link will disappear in the next time window (Figure. 2.8). This problem is more complex than adding link problem as it might need more information to perform prediction task.

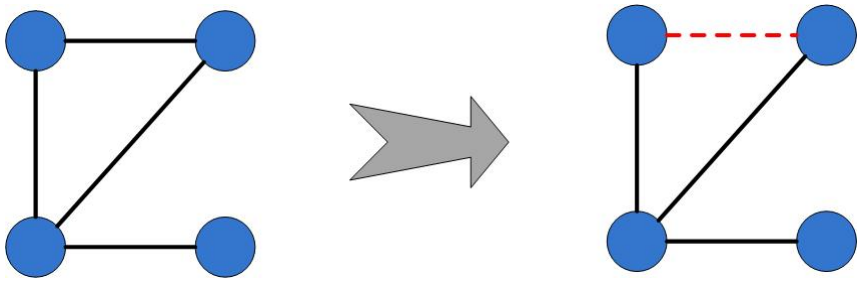


Figure 2.8: Removing Link Prediction

Adding and Removing Links

This problem is a combination of previous two problems (Figure. 2.9). It means that one simultaneously predicts both creation and removal of different links that will tke place from one time window to another one.

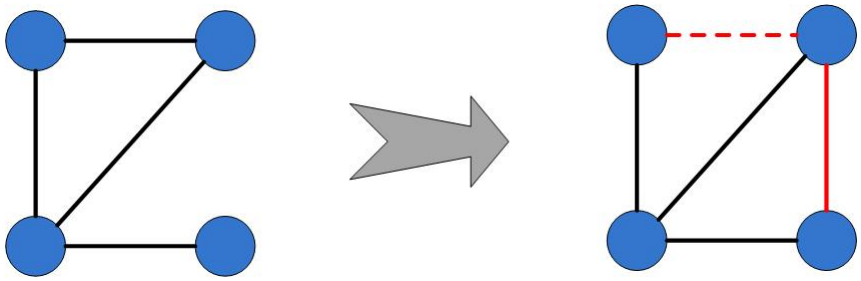


Figure 2.9: Adding and Removing Link Prediction

There are several link prediction methods and they are presented below.

2.4.1 Common Neighbours

This prediction method is on the assumption that two nodes with many common neighbours will be connected in the future. It is also an intuitive method which ranks the nodes pairs according to the number of common neighbours. As another basic prediction method, it is also usually used as a baseline to judge the performance of other methods. It is usually the best performing method among the basic prediction methods[26; 29; 31; 73].

$$|\Gamma(i) \cap \Gamma(j)| \tag{2.5}$$

Where $\Gamma(i)$ and $\Gamma(j)$ represent the set of neighbours of node i and node j.

2.4.2 Jaccard's Coefficient

The Jaccard's Coefficient, also known as Jaccard index or Jaccard similarity coefficient, is a statistic measure used for comparing similarity of sample sets. It is usually denoted as $J(x, y)$ where x and y represent two different nodes in a network. In link prediction, all the neighbours of a node are treated as a set and the prediction is done by computing and ranking the similarity of the

neighbour set of each node pair. This method is based on Common Neighbours method and its complexity is also $O(nk^2)$. The mathematical expression of this method is as follows [29]:

$$\left|\frac{\Gamma(i) \cap \Gamma(j)}{\Gamma(i) \cup \Gamma(j)}\right| \tag{2.6}$$

2.4.3 Adamic/Adar Index

It was initially designed to measure the relation between personal home pages. As shown in equation 2.7, the more friends z has, the lower score $AA(i, j)$ will be. A common neighbour of a pair of nodes with few connections contributes more to the similarity score between the two nodes than common neighbour with a lot of connections. In real world, the phenomenon is like this, if a common acquaintance of two people has more friends, then it is less likely that he/she introduces the two people to each other. It shows good result in predicting the friendship in personal homepages and Wikipedia Collaboration Graph, but in the experiment of predicting author collaboration, it shows a poor prediction performance [25].

$$\sum_{z \in \Gamma(i) \bigcap \Gamma(j)} \frac{1}{\log k_z} \tag{2.7}$$

Where i and j is a pair of nodes. z is one of the common neighbours of i and j. $\Gamma(i), \Gamma(j), \Gamma(z)$ is the set of neighbours of node i, j and z respectively.

2.4.4 Sørensen Index

This index [74] is designed for comparing the similarity of two samples and originally used to analysis plant sociology. It is defined as:

$$\frac{2|\Gamma(i) \bigcap \Gamma(j)|}{k_i + k_j} \tag{2.8}$$

Where k_i and k_j stands for the degree of node x and node y respectively.

2.4.5 Katz$_\beta$

The relationship between two nodes is represented by the links between them, so it is reasonable to take all the path lengths into consideration when doing a link prediction. Katz$_\beta$ is one of the methods based on this paradigm. According to equation 2.9, the number of paths between node i and node j with length l (written as $|paths_{ij}^{\langle l \rangle}|$) are calculated and then multiplied by a factor β^l. By summing up all the results for two nodes with path length from 1 to ∞, a prediction score of the pair of nodes (i, j) is obtained. The parameter β, as shown in equation 2.9, is used to adjust the weight of path with different length. For instance, when an extremely small β is chosen, the longer path will contribute less to the result as β^l could be very small as the path length l is getting larger. Thus, the result will be similar to the common neighbours. A, A^2 and A^3 denote adjacency matrices about the nodes having 1 length, 2 length and 3 length distances, respectively. It is one of the well performing prediction methods in many experiments and it takes the contribution of paths of arbitrary length into account. [75]

$$\sum_{l=1}^{\infty} \beta^l \cdot |paths_{ij}^{\langle l \rangle}| = \beta A + \beta^2 A^2 + \beta^3 A^3 + ... \tag{2.9}$$

2.4.6 Cosine Similarity

The idea of this method comes from the dot product of two vectors. It is often used to compare documents in text mining tasks. In addition, it is used to measure cohesion within clusters in the field of data mining. However, the performance in network prediction is not clearly mentioned in papers thus more experiments are needed to see how it works [31].

$$\frac{|\Gamma(i) \bigcap \Gamma(j)|}{\sqrt{k_i * k_j}} \tag{2.10}$$

2.4.7 Preferential Attachment

Due to the assumption that the node with high degree is more likely to get new links [76], preferential attachment can be used as prediction algorithm. In link prediction problem, the degrees of both nodes from a given pair need to be considered for the prediction score calculation. This can be easily found from the mathematical expression bellow. The score can be calculated by multiplying the degree of both nodes. Same as common neighbours and graph distance, this is also a basic prediction method which is usually used as a baseline to measure the performance of other prediction methods. In [29], preferential attachment is not the best performing method for almost all the networks. However, in [77], author points out that preferential attachment has a strong correlationship with gini coefficient. It is defined in equation 2.11. k_i and k_j are the degree of nodes i and j.

$$k_i * k_j \tag{2.11}$$

2.4.8 Resource Allocation Index

This index is motivated by the resource allocation dynamics on complex networks [78]. Consider a pair of nodes, i and j, which are not directly connected. The node i can send some resource to j, with their common neighbors playing the role of transmitters. In the simplest case, it assumes that each transmitter has a unit of resource, and will equally distribute it to all its neighbours. The similarity between i and j can be defined as the amount of resource j received from i. Resource allocation performed very well comparing to other methods in [31]. AA and RA have very close prediction results for the networks with small average degrees, but RA performs better for the networks with high average degrees [79].

$$RA(i,j) = \sum_{z \in \Gamma(i) \bigcap \Gamma(j)} \frac{1}{k_z} \tag{2.12}$$

2.4.9 Hub Promoted Index

HPI is proposed for analysis of metabolic networks as shown in [80]. The property of this index is that the links adjacent to hubs are likely to obtain a higher similarity score. It is expressed as:

$$\frac{|\Gamma(i) \bigcap \Gamma(j)|}{min\{k_i, k_j\}} \tag{2.13}$$

2.4.10 Hub Depressed Index

Approach that uses the idea of hub in totally different manner than HPI is Hub Depressed Index (HDI). It gives links adjacent to hub a lower score. It is defined as

$$\frac{|\Gamma(i) \bigcap \Gamma(j)|}{max\{k_i, k_j\}} \tag{2.14}$$

2.4.11 Leicht-Holme-Newman Index

Leicht-Holme-Newman Index (LHNI) [81] was proposed to quantify the similarity of nodes in networks. It is based on the concept that two nodes are similar if their immediate neighbours in the network are themselves similar. It is defined as:

$$\frac{|\Gamma(i) \bigcap \Gamma(j)|}{k_i * k_j} \tag{2.15}$$

In our work, we will use these prediction methods for study and the measurement of their performance will be described in Section 3.5.

2.4.12 Discussion

As discussed at the beginning of this section, there are three types of problems in network prediction. They are link prediction, node prediction, link & node prediction. Here we reviewed the first one in detail.

To sum up, most of the classic link prediction methods only focus on the topology information of the network. The common character of them is the assumption that all the nodes are homogeneous. These classic methods have better prediction performance than pure random prediction [29; 31], but none of the methods shows an outstanding performance. All these methods are applicable on undirected networks. For directed network, they could also be applied on the in-coming and out-going links. With the development of World Wide Web and online social networks, more complex networks with richer information are available and more tools are developed to help researcher work easier (i.e. NetworkX [82]). All of this inspire many researchers who put effort into developing new approaches to link prediction problem. On the premise of this, the link prediction problem could be reconsidered from more angles such as network structure information like community.

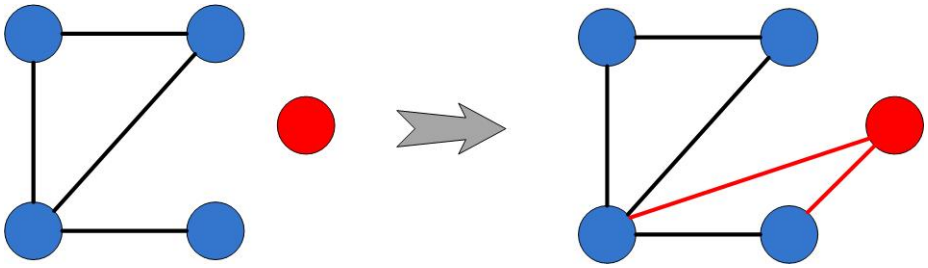


Figure 2.10: Adding Node Prediction

2.5 Node Prediction

The changes in complex networks include both changing number of links and nodes. Thus, node prediction is another research topic is researched. Similar to link prediction, nodes prediction also contain three types of problems.

Adding Nodes

Obviously, when a new node is added to the network, new links will be formed. Different with adding link problem, adding nodes problem cares more about to which existing nodes the new one will be connected (Fig.2.10).

Removing Nodes

Removing nodes concentrates on the prediction of the node disappearance (Fig. 2.11). As a consequence, removing nodes always leads to removing links that were connected to these nodes.

Adding and Removing Nodes

Adding and removing nodes problem, as shown in Fig.2.12, is the combination of previous two problems. This is also the most complicated problem in node prediction task.

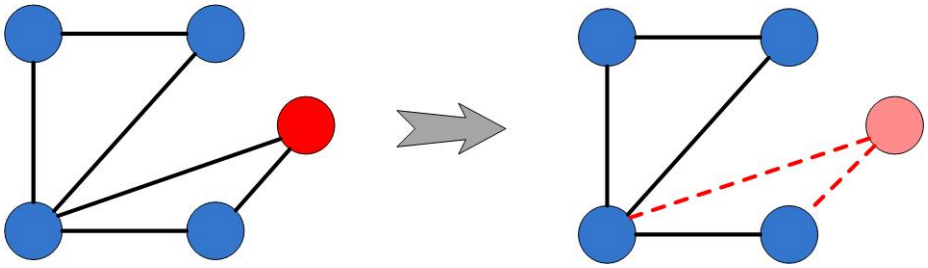


Figure 2.11: Addving Node Prediction

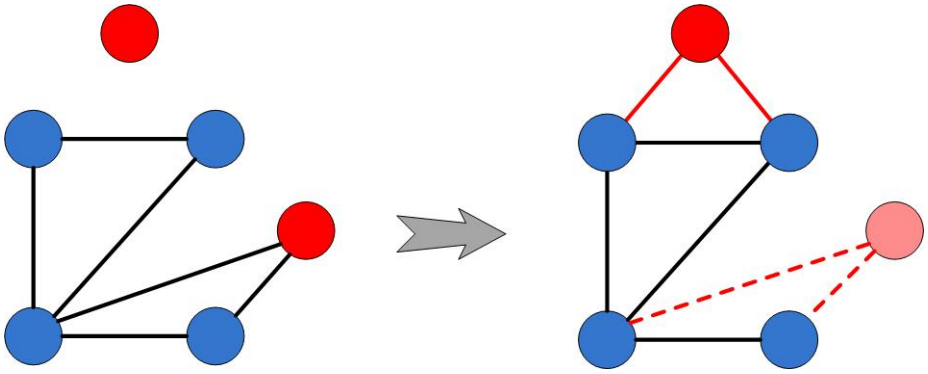


Figure 2.12: Addving and Removing Node Prediction

2.5.1 Random Growth Model

Nodes prediction solve the problem that to which the new coming node connected. The prediction is also based on probabilities (i.e. the new coming nodes have a higher probability connecting to A than B). There is no intuitive approach to do node prediction by only considering topology information of network. A random growth model try to simulate the growth of complex network. The new coming node has equal probability to connect to the old nodes. The equal probability means useless from the perspective of prediction. However, the probability is not same for all the nodes, it is positive correlation to the degree of old nodes which reflect the phenomenon of preferential attachment. Thus, a model-driven node prediction method could be introduced by ranking the degree of all the old nodes. Higher ranked nodes has higher connection priority which is a kind of prediction.

Barabási-Albert model

One of the most popular random growth models is BA model introduced in section 2.3.4. It was initially designed to create scale-free network. The model grows network by adding nodes with fix number of degree. New node is more likely to connect to existing node with higher degree. The result network generated by BA model follows a power-law degree distribution.

Fitness Model

The BA model has its limit. The new coming nodes generally always has a lower connection priority than old nodes due to the "rich and richer" effect. The accuracy of prediction based on the model thus might be affected. In [83], Bianconi and Barabási introduced fitness model which is a variant to the BA model. The basic idea is "fit get richer". A time independent fitness factor will assign to each node. The probability of connection of new node then calculated with involve of fitness parameter. The prediction then become the problem of how to define the fitness parameter. In real world networks, the fitness could be defined according to the attribute information of nodes or the topology information and the prediction process is similar to the method discussed above.

2.5.2 Discussion

The research on complex network has been done for a long time. However, more progresses has been made in recent decades years as there are more data available for the study. Currently, most of the research still focus on the link prediction. There are limit literature on node prediction and this is also the

reason why the other two (removing node, adding and removing) problems are not included in detail.

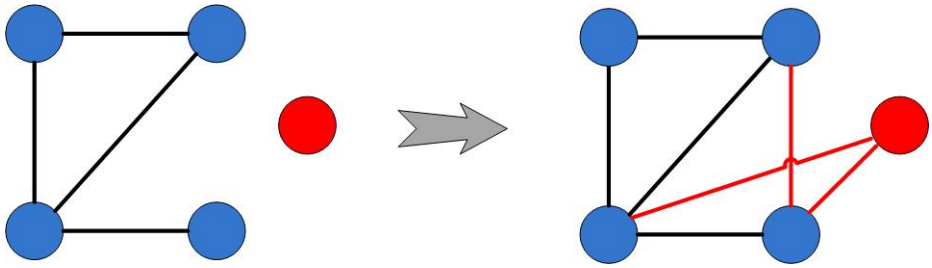


Figure 2.13: Addving Link & Node Prediction

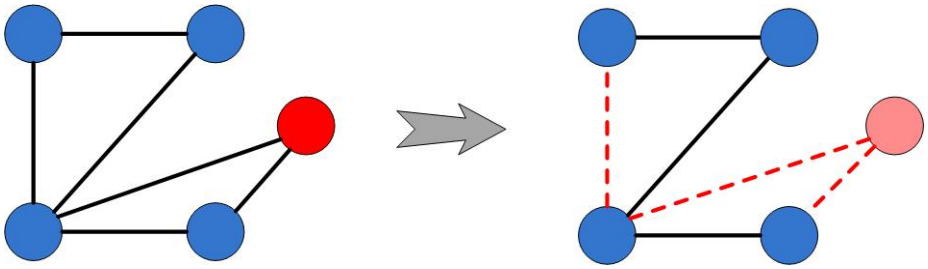


Figure 2.14: Removing Link & Node Prediction

2.6 Links & Nodes Prediction

The ultimate aim of complex network prediction is to predict both the changes in number of links and nodes at the same time. However, most current researches still stay at the level of link prediction. Nevertheless, it is very important to have a clear and coherent overview of the whole problem space.

Adding Links & Nodes

This problem considers the appearance of nodes and links in the future at same time (Fig. 2.13).

Removing Links & Nodes

Removing links and node problem considers the disappearance of both links and nodes in same time (Fig. 2.14)

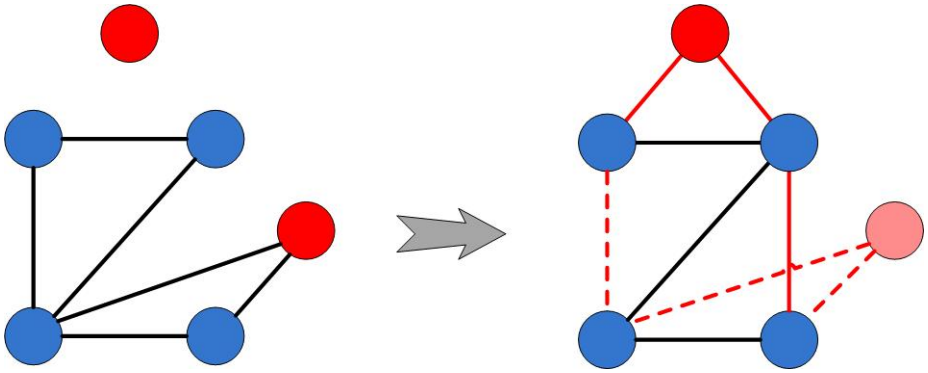


Figure 2.15: Adding and Removing Link & Node Prediction

Adding and Removing Links & Nodes

This is the ultimate aim of complex network prediction. The ideal prediction method could predict the change of network includes both adding and removing of both links and nodes (Fig.2.15). Technically this is also the most challenging one in prediction problems.

2.6.1 Discussion

Links and nodes prediction at the same time is the most desirable method for complex network prediction and it is really a big challenge. Combining the sophisticated node and link prediction methods could be a way. However, so far, there are no well performing link prediction methods and existing node prediction methods are limited and lots of study still needs to be done.

Chapter 3

Methodology

This chapter introduces the methodology that is followed in this project. The main goal of the project is to propose a new prediction method for large social networks which could provide a better prediction accuracy than existing approaches. To achieve the research goal and answer the research questions, the hypotheses must be verified. The whole research process from problem formulation to final conclusions is presented in Figure 3.1.

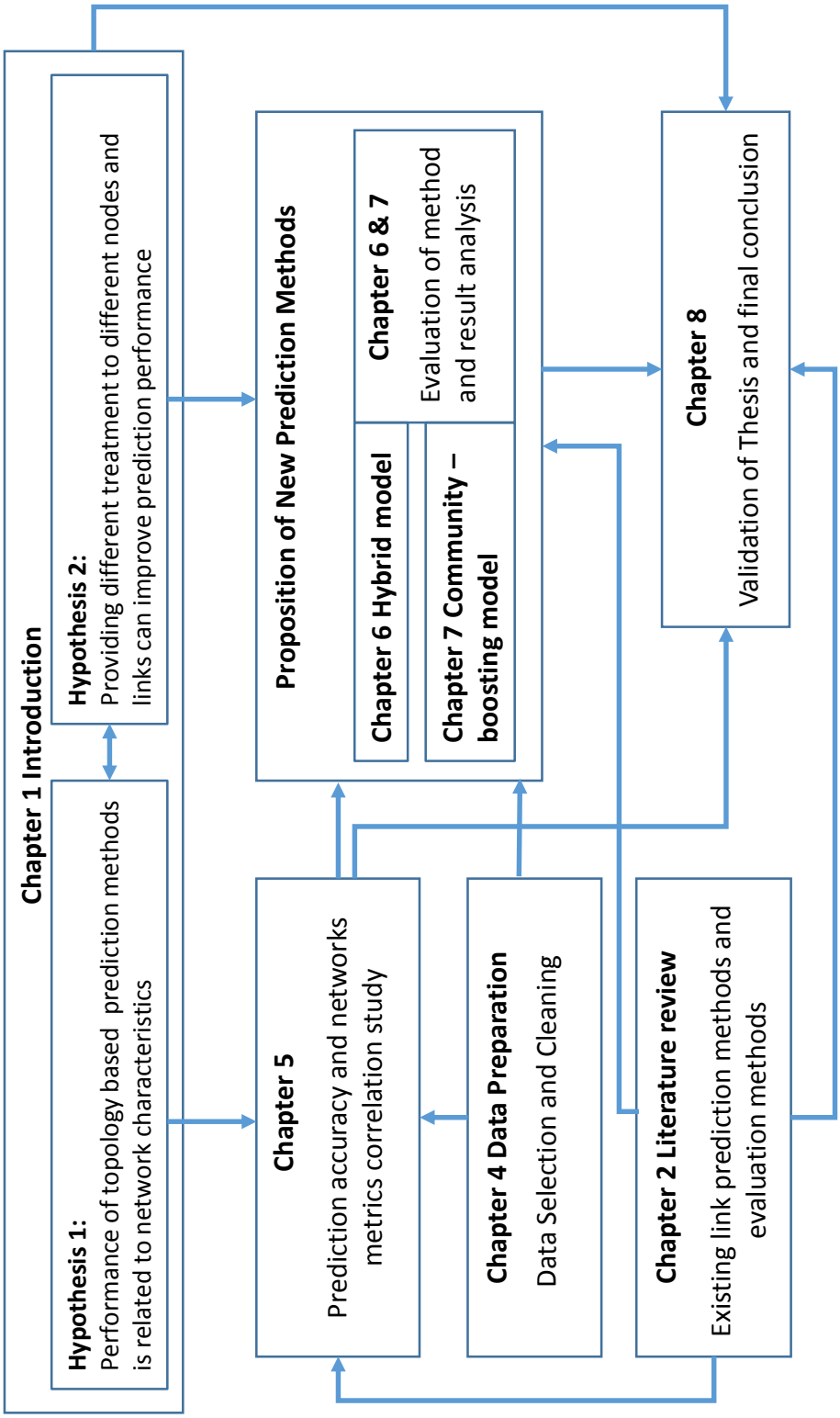


Figure 3.1: Research Process

The whole research will be divided into the following stages:

1. Based on the existing research result, formalize the two hypothesis for the research (Chapter 1).

2. Review the existing link prediction methods and evaluation methods (Chapter 2).

3. Select and prepare social network datasets to test the proposed link prediction method (Chapter 4).

4. Perform the study of correlation between network metrics and the topology based link prediction methods (Chapter 5).

5. Propose two link prediction models based on the second hypothesis, Hybrid model (Chapter 6) and Community Boosting model (Chapter 7). Evaluate and analysis the result.

6. Summarize and validate of my study and thesis. Make the final conclusion for my PhD work (Chapter 8 & 9).

3.1 Literature Review

Before proposing and developing our link prediction model, recent research achievements are reviewed to help gain the understanding of the work background. They are reviewed from three angles: Network Metrics, Network Models and Structure-based Network Prediction Methods. The detailed research works are then designed, implemented and performed on top of the review.

3.2 Data Selection and Cleaning

The target of this study is online social network. By online social network, we refer to the network that is generated by a group of people with any forms of contacts or interactions between them via Internet. Thus, the internet based social networks are the most suitable data for this research. The example could be online friendship network, email network, communication network. etc. Another principle for data selection is that timestamp of network link formation must be included in dataset. Because the timestamp information reflects the evolvement of dynamic networks.

With these assumptions, we selected eight different online social networks which will be introduced in detail in Chapter 4.1. To ensure our study is applicable to general social networks, we did not filter the social networks we selected with specific criterion. Meanwhile, there are also many theoretical network model which also reflect network evolutions (e.g. Barabási-Albert model). They enable network growing with different restrictions which we believe cannot describe real world network evolution, thus, we did not use them in our work for model testing. The purpose of this research is for general social network, but we have to point out that the limitation of our work still exist. The hybrid model only applicable to networks with long evolution time scales while the community bridge boosting model can be used only on networks with multi-communities.

In this study, we use the network topology information to predict new links in network. Nodes or small network components that could not reveal whole network topology characteristics would introduce noises in the experiment. To reduce such noises generated by isolated nodes and isolated small cliques, we performed data cleaning for all the selected datasets by extracting the giant component for study. Details of data selection and cleaning are presented in

Chapter 4.

3.3 Correlation Analysis

The Hypothesis 1 is to study the correlation between the performance of link prediction methods and network metrics. The work flowchart of this study as introduced in detail in Chapter 5 is shown in Figure 3.2. Six networks are selected for this study include: Enron Email network, Facebook wall post network, Flickr friendship network, PWr Email network, UC Irvine message network and Youtube Friendship network. They are all internet based social networks. The steps are as follows:

1. For each of the network, calculated the six network metrics as shown in Figure 3.2 for further correlation study.

2. Applying the ten selected prediction methods for each network. The accuracy are measured using AUC.

3. The correlation coefficient between the AUC results and network metrics are calculated and then analysed.

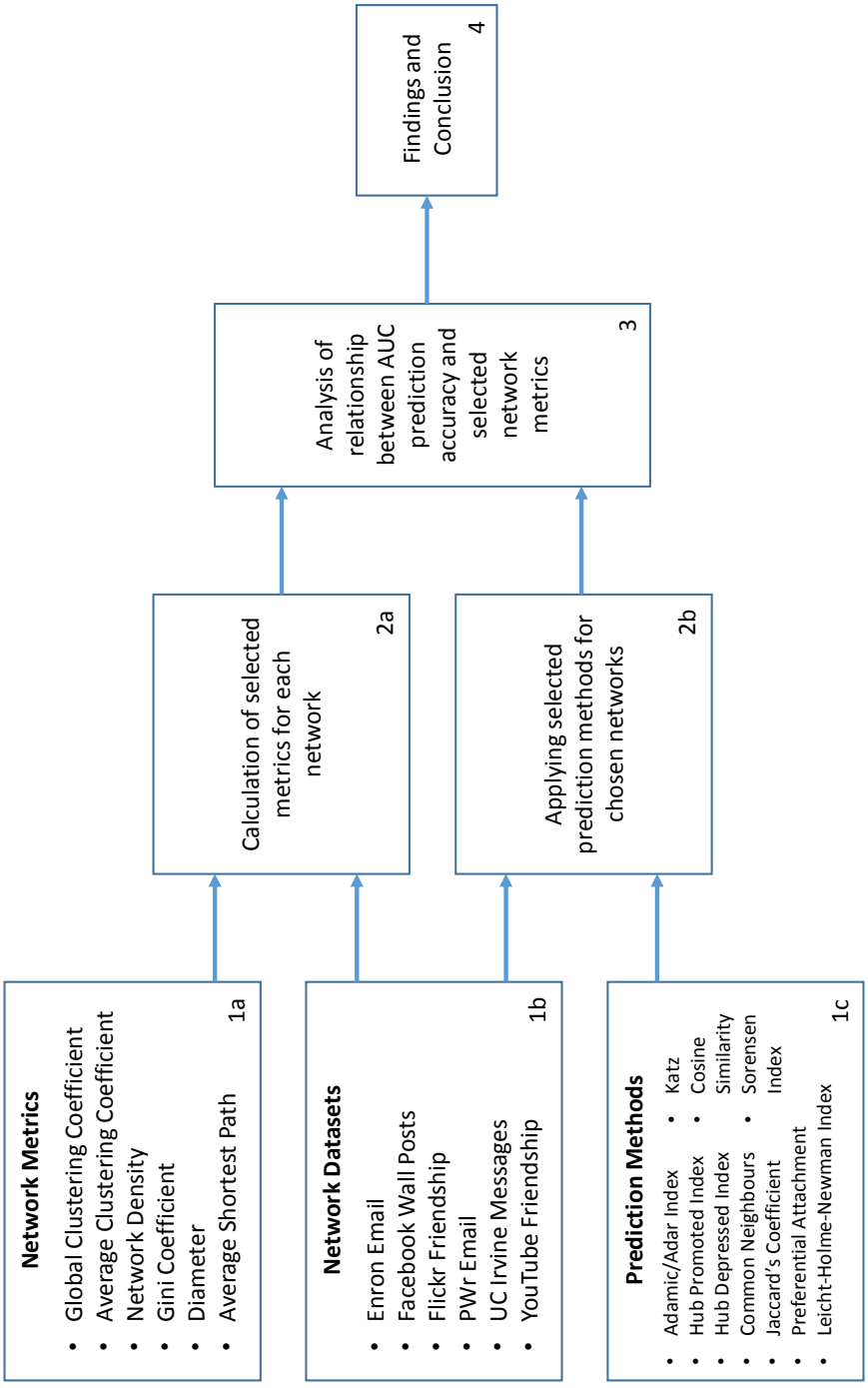


Figure 3.2: Correlation between link prediction accuracy and network metrics

3.4 Proposition of New Prediction Methods

Based on the prediction methods and networks metrics correlation study, we concluded that different networks evolve following different rules. The rules can change over time so prediction model should be able to self-adapt to the changing rules. This means that different methods will work well for different networks. Moreover, different prediction methods can result in varying prediction accuracy for different parts of networks. Thus, in this project we propose two approaches (i) hybrid method that is an ensemble of different prediction methods and (ii) community-boosting approach where different parts of networks will be predicted using predictor that is personalised to a specific region of a network.

3.4.1 Hybrid Model

By assuming network evolvement following different rules, we proposed the hybrid model. This model linearly combined several different link prediction methods. By solving a optimization problem, we obtain the best weight for each method in the combination which we refer to the network evolution rule. The work flow of this research is shown in Figure 3.3 and describe in more detail in Chapter 6. The model is capable to work with two scenarios as following.

Sliding Window

Sliding window scenario slides the window without remembering the historical network information. The prediction is done only based on the network information within a given window. This scenario focuses only on the recently formed links.

Growing Window

In growing window scenario, the window grows so that all the historical network information is used for prediction. The network information will be richer as the window grows thus this scenario focuses on all the historical data.

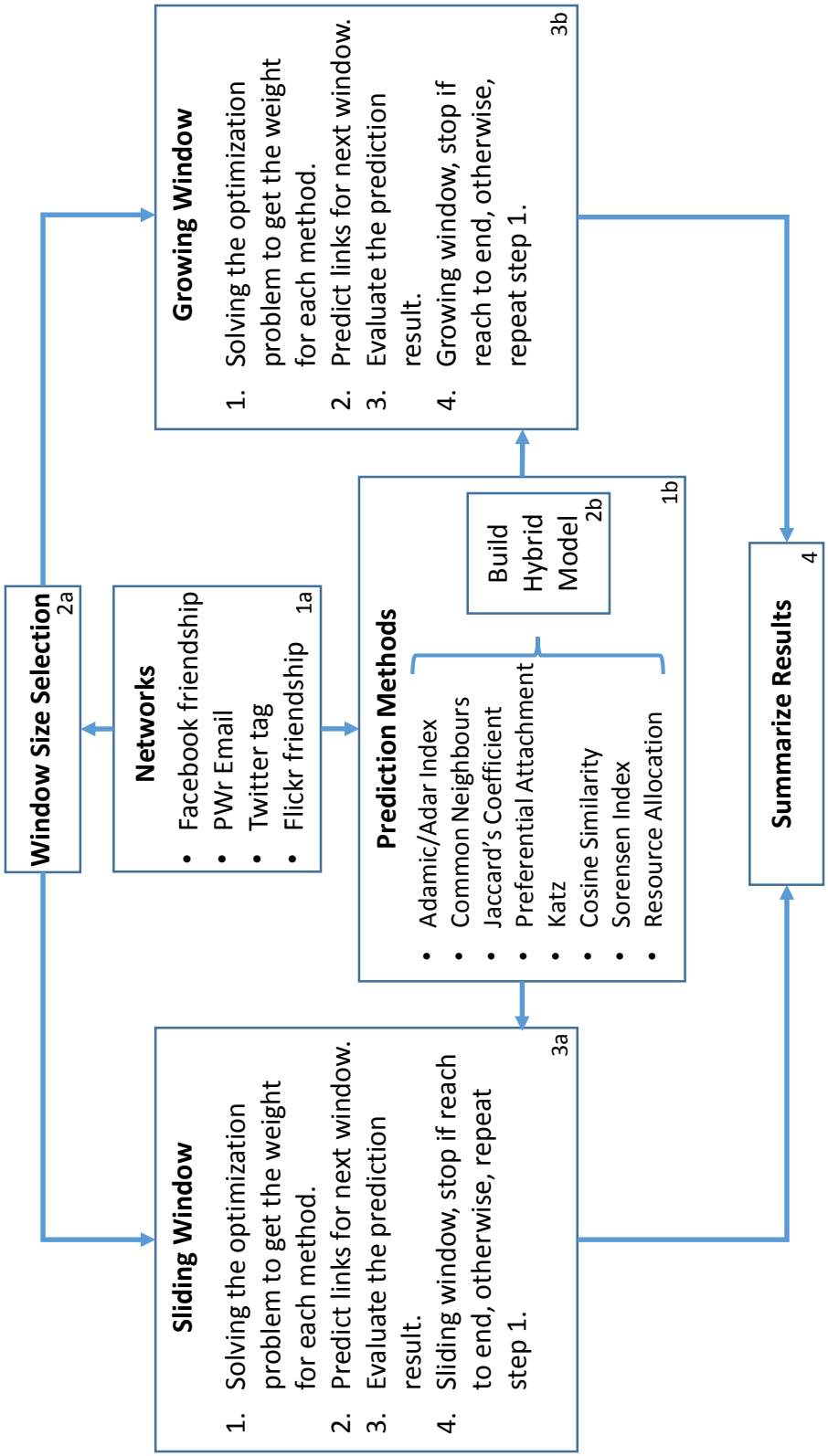


Figure 3.3: Hybrid Prediction Model Work Flow

Window Size Selection

As we are using sliding and growing windows for our model, two important questions need to be solved before the experiment are the window size and the window moving step size. We select one month and one week as our window step size according to human social life cycle in real life. The window step size is used to grow or sliding the windows for each step. With the selected window size, we calculate the optimized window size with the method proposed in [84].

3.4.2 Community Bridge Boosting Prediction Model

To providing different treatment to nodes, we proposed Community Bridge Boosting Prediction Model (CBBPM). We define and classify network nodes as community bridge node based on their degree and links position in network communities. The similarity score that calculated from the selected prediction methods is then boosted for predicting new links. Its work flow is shown in Figure 3.4 and introduced in Chapter 7.

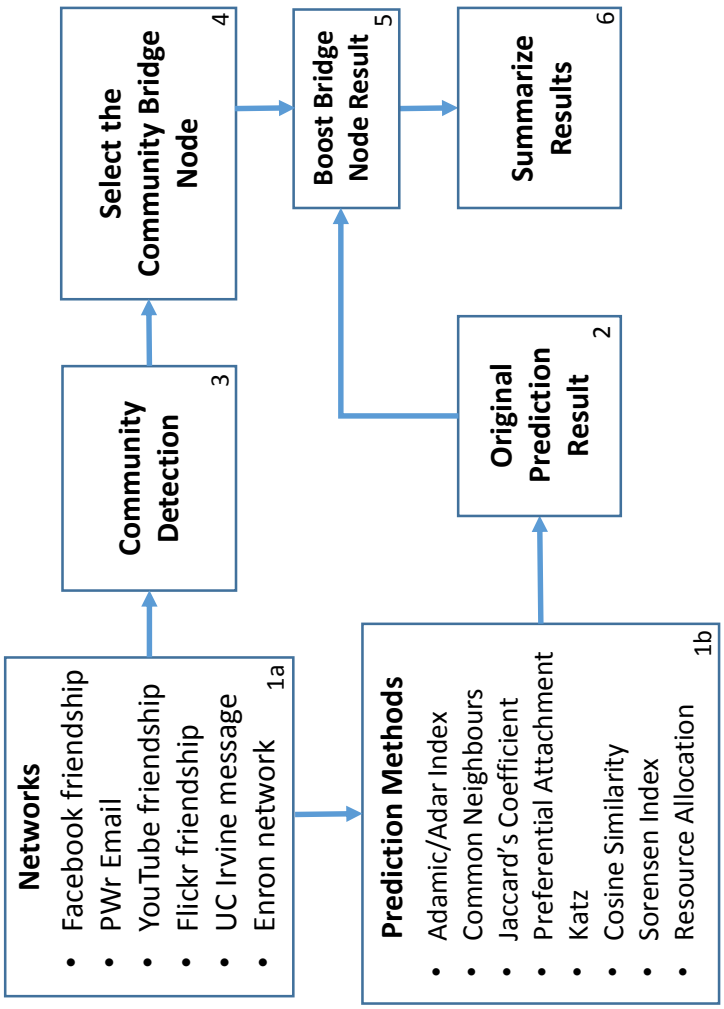


Figure 3.4: Community Bridge Boosting Prediction Model Work Flow

3.5 Evaluation of Link Prediction Methods

The new methods must be verified in order to check its correctness and accuracy. Following are some common methods to evaluate the performance of method. We will compare our method with other methods using those evaluation metrics.

In order to measure the performance of a prediction method, we need to use historical network data. Link prediction is a time related activity, therefore, we should use time–stamped dataset and according to the time stamp, separate the data into two sets, $G_{t,t_1}(N, L_1)$ as training set for prediction methods and $G_{t_1,t_2}(N, L_2)$ as unknown future network for testing where $t < t_1 < t_2$. Those two networks must consist of the same set of nodes V. The number of possible links that is denoted by U is $|N| * (|N| - 1)/2$. The link prediction method, in principle, provides a similarity score for each non existing links $(U - L_1)$ and for most methods, a higher score means higher likelihood that the link will appear in the future. Final prediction is done by ordering this score list and selecting top N links with the highest score.

3.5.1 Precision

This is a basic measure method to quantify the prediction accuracy. For a predicted link set L, there are some links that do appear in the future which means the prediction is correct and is denoted as L_r. The remaining set $L - L_r$ represents links that did not appear. Precision is the ratio of the number of correct predictions L_r and the number of all predictions L [31].

$$Precision = \frac{L_r}{L} \tag{3.1}$$

3.5.2 Recall

Recall is statistical measure of the performance of a binary classification test. Prediction method evaluation, as a binary query of whether the prediction is right or wrong, is a kind of problem that the measure is designed for. The measure is composed of three variables [85]:

True Positive (TP): The number of correctly predicted relationships. In network prediction problem, it means the number of links or nodes that in the prediction set do appear in the future.

False Positive (FP): The number of objects that do not appear in predictions set. Here, it is the number of links or nodes that not in the prediction set that do not appear in the future.

False Negative (FN): The number of objects that are not in the prediction set but appear in the future. Here in the research, it is the number of links or nodes that do not appear in the prediction set but appear in the future.

And then the Recall is defined as follows:

$$Recall = \frac{TP}{TP + FN} \tag{3.2}$$

In this context, the Precision that defined in Section 3.5.1 can also be written as:

$$Precision = \frac{TP}{TP + FP} \tag{3.3}$$

3.5.3 Area Under the Receiver-Operating Characteristic (AUC)

In our work, AUC is used for quantifying the accuracy of prediction method. It is the area under the receiver operating characteristic curve [86]. In the context of network link prediction, if L_1 represent the links of network snapshot and L_2 represent the links of network snapshot later, U as the all possible links between the nodes, then AUC can be interpreted as the probability that a randomly chosen missing links $(L_1 \cup L_2 - L_1)$ is given a higher similarity score than a randomly chosen pair of unconnected links $(U - (L_1 \cup L_2))$ [87]. The algorithmic implementation of AUC follows the approach in [31]. It is calculated as:

$$\frac{l' + 0.5l''}{l} \tag{3.4}$$

Where l is the number of times that we randomly pick a pair of links from missing links set and unconnected links set; l' is the number of times that the missing link got a higher score than unconnected link while l'' is the number of times when they are equal. The AUC value will be 0.5 if the score are generated from and independent and identical distribution. Thus, the degree to which the AUC exceeds 0.5 indicates how much better the predictions when compared to predict by chance.

3.6 Conclusion and Future Work

To close the research circle, the last Chapter reviews the research question and our hypotheses that introduced in Chapter 1. The conclusions and finds are summarized with potential future works.

Chapter 4

Data Preparation

This chapter introduces datasets we used for the experiments in this thesis. Below, the description about how each dataset for each experiment was cleaned and processed is provided.

4.1 Datasets Selection

In total, we select eight Internet based social networks. To make sure the datasets fit the research problem and experiments requirements, they are selected based on two principles:

1. The network must be collected from the World Wide Web, or in another words, it is online based networks.

2. For each link in the network, their formation time must be available.

All the original dataset information are shown in Table 4.1. All the datasets apart from PWr Email communication network are from Koblenz Network Collection (KONECT [88]). The PWr Email communication network is the internal email communication network from Wrocław University of Technology [89]. The nodes in this network represents a person and link between nodes

reflects email communication between them. Similarly, we also select Enron Email communication network from [90]. It is the Email network among employees of Enron before its bankruptcy. Another similar network we selecte is the UC Irvine message network [91]. It contains messages send between the users of an online community of students from the University of California, Irvine. A node represents a user and a link represents sent message. Multiple links denote multiple messages. The Facebook wall post network is the wall posts from the Facebook New Orleans networks [92]. In this network, a node stands for a user while a link stands for a message wall posts between two users. Flickr network is about the users and their online friendship connections on the website. It is collected by taking a snapshot of the network on November 2nd, 2006 and record it daily until December 3rd, 2006, and then again daily between February 3rd, 2007 and May 18th, 2007 [93]. The YouTube Friendship network depict the user followship on the website. It was originally collected between December 10th, 2006 and January 15th, 2007, and again daily between February 8th, 2007 and July 23rd, 2007 [94]. We also select another friendship network of Facebook from [92] for study. The last network we use for experiments is Twitter network [95; 96]. It was collected in one week between 2009/10/13 to 2009/10/20. Same as other networks, each node is a user. Each link in this Twitter network denotes a '@username' tweet sent from user A to another user, user B who is using the username.

4.2 Data processing

In this research, to avoid over complicating the study as well as taking into account the fact that most of our benchmarking methods work for undirected and unweighted networks, we treat all the selected networks as undirected

Dataset Name	Time Range	No. of Nodes	No. of Edges
Enron E-mail	1998/11 - 2002/07	87,273	1,148,072
Facebook Wall Posts	2008/01 - 2009/01	63,731	1,269,502
Flickr Friendship	2006/11 - 2007/05	2,302,925	33,140,018
PWr E-mail	2008/11 - 2009/05	14,316	49,950
UC Irvine Messages	2004/03 - 2004/10	1,899	59,835
YouTube Friendship	2006/12 - 2007/07	3,223,589	12,223,774
Facebook Friendship	2007/01 - 2007/06	8,564	33,950
Twitter	2009/10 - 2009/10	2,919,613	12,887,063

Table 4.1: Original Dataset Information

unweighted networks. The following sections introduce how the datasets have been processed for each experiment.

4.2.1 Prediction Accuracy and Network Metrics Correlation Study

In order to verify the Hypothesis 1 (The performance of topology based network prediction methods and the characteristics of the networks are correlated), we conduct the correlation coefficient study between Prediction Accuracy of selected methods and selected Network Metrics. We select six networks out of the eight with the following two reasons: We prefer to using networks with longer time dimension (in this experiment we choose networks for which we captured evolution for more than 6 months) because it contains more new links. So we select all the six networks that meet the time dimension requirement from the candidate networks.

Once the datasets have been selected, we processed them with following steps:

1. **Select data samples**. For each dataset, we first randomly select 6000 - 8000 user records (8000 samples is selected due to the calculation capacity. As for some dense networks, 8000 nodes is also too big, so we choose 6000) from the original dataset as the sample user and all connections of selected nodes. As UC Irvine Messages only contains 1899 users, so we leave it as it is. The specific sample numbers are shown in Table 4.2.

2. **Split the data into training and testing sets**. Prediction in a time series problem means the dataset should be divided into train and test sets based on time stamps available. As the dataset of Flickr and YouTube are collected by taking snapshot of the network which is different from other four datasets, we take the first day snapshot as the training set and the remaining data as the test set. The other four networks are split according to the time scale with a ratio approximate training time : test time = 80% : 20% as shown in Table 4.2.

3. **Extract giant component**. Dividing data into training and testing sets can cause the isolation of some nodes or cliques. This, in turn, generates noise for measuring the accuracy of prediction methods as the methods we selected cannot predict unconnected nodes. To eliminate the impact of this noise, we extract the giant component from training dataset as our final training set $G_{t,t_1}(N, L_1)$. The final test set $G_{t_1,t_2}(N, L_2)$ is obtained by extracting the network with all the nodes existing in $G_{t,t_1}(N, L_1)$ from the original test set obtained from step 2. For nodes existing in the final training set but not present in the original test set, we just keep and leave them isolated in the final test set.

After all, we get the train set $G_{t,t_1}(N, L_1)$ and test set $G_{t_1,t_2}(N, L_2)$ that both have the same nodes set.

Dataset Name	Train Time Range	Test Time Range	Sample Nodes	Final Nodes
Enron E-mail	1998/11 - 2001/12	2002/01 - 2002/07	8000	5208
Facebook Wall Posts	2008/01 - 2008/11	2008/12 - 2009/01	8000	5784
Flickr Friendiship	Snapshot on 2006/11/02	2006/11/03 - 2006/12/03& 2007/02/03 - 2007/05/18	6000	5949
PWr E-mail	2008/11 - 2009/04	2009/04 - 2009/05	8000	6335
UC Irvine Messages	2004/03 - 2004/08	2004/08 - 2004/10	1899	1666
YouTube Friendship	Snapshot on 2006/12/10	2006/12/11 - 2007/01/15& 2007/02/08 - 2007/07/23	6000	6000

Table 4.2: Prediction Accuracy and Network Metrics Study Experiment Datasets

4.2.2 Hybrid Prediction Model

Based on Hypothesis 2 (As network are dynamic, the performance of prediction can be improved by providing different treatment to nodes and links), we proposed a hybrid prediction model which provide different treatment for different links. The model details is introduced in Chapter 6. To test the hybrid model, we select four real world network datasets, the Facebook friendship network [92], the internal email communication network from Wrocław University of Technology, the Flickr following network[93] and Twitter Hashtag Network [95] . Table 4.1 shows the network information. In Facebook network, each node represents a user and the link between two nodes means they are friend. For the Email communication networks, nodes are users and the link between the two nodes, A and B, means an email was sent between them, either from A to B or B to A. In the Twitter hashtag network, nodes are users and links represent that the two nodes communicate with each other using hashtag in their Twitter content. The Flickr network contains the network where nodes are users and links are the following relationship on the website. Each link in both datasets has a time stamp which records the time when the link was formed. We take all datasets as binary, directed, un-weighted networks.

As shown in Table 4.1, there are 14,316 nodes in the PWr Email communication network. However, we find that among these users, only 6,884 users sent email at least once. Rest of the accounts only receive emails without any other activities. Similarly, in Twitter network, we can find that among the 2,919,613 nodes, only 3,172 users have both outgoing and incoming relationships. We treat nodes only have incoming links as inactive user and thus removed all of these nodes with no outgoing link so that only active users who sent at least one email or has at least one tweet are kept for the experiment. We also removed from the dataset isolated small cliques as they are not connected with

majority of nodes which would bring in noise when perform link prediction. This is achieved by extracting the giant component from the four networks. Table 4.3 shows the network information after the cleansing process.

Name	Time Range	Nodes	Links
Facebook	2007-01-01 to 2007-06-30	7,446	27,140
PWr Email	2008-11-25 to 2009-05-25	6,059	27,640
Twitter	2009-10-13 to 2009-10-20	1,564	2,376
Flickr	2008-11-25 to 2009-05-25	5,949	408,086

Table 4.3: Giant Component Network Information

Window Size and Window Step Size

The hybrid model is capable to predict in two scenarios indroduced in Section 3.4.1, Sliding Window and Growing Window. Before implementing the two scenarios, there are two important questions need to be answered: What is the best windows size and window step size to be used for hybrid model testing?

Taking into account human social life cycle, we select two window sizes for our experiments – week and month. A week is defined by 7 consecutive days and a month is defined by 28 consecutive days (4 weeks). Another issue is the size of the step by which we slide or grow the window. To address this, we used method introduced in [84]. Authors claim that by choosing window size in a way that the properties of a network within each window are as close as possible to the characteristics of the global network, the link prediction accuracy can be increased. With considering four characteristics, node degree distribution divergence, the shortest path length distribution divergence, the clustering coefficient divergence and the betweenness centrality divergence introduced in [84], we obtained the optimal step size for all networks as shown in Table 4.4. The selected step size applies to both sliding and growing window scenarios. The Twitter network only contains nodes and links in one week, we thus select one day for both window size and window step for our experiments thus it is not stated in Table 4.4.

Neworks	Monthly Window Step	Weekly Window Step
Facebook	14 days	6 days
PWr	28 days	5 days
Flickr	21 days	7 days

Table 4.4: Optimal Window Step Size

4.2.3 Community Bridge Boosting Prediction Model

We also proposed Community Bridge Boosting Prediction Model (CBBPM) by providing different treatment to network nodes (in Chapter 7). This model is based on community information thus the network used for this study cannot be too sparse. Otherwise, it is hard to detect meaningful communities with large number of nodes. So we need to select networks that have a long time scale. This principle is same as the one for Prediction Accuracy and Network Metrics Correlation Study introduced in Section 4.2.1. Considering the reusability of datasets, we decided to select the same train and test networks as stated in Table 4.2 to test CBBPM. These networks are: Enron Email Network, Facebook Wall post Network, Flickr Network, PWr Email Network, UC Irvine Message Network and YouTube Network. They were pre-processed in the same way as for the Prediction Accuracy and Network Metrics Correlation study.

Chapter 5

Prediction Accuracy and Network Metrics Study

In this chapter, the correlation coefficient study between structure based link prediction accuracy and network metrics is introduced. The experiments results are provided and analysed. The last section summarized all the conclusions and findings.

5.1 Study Background and Motivation

In the network prediction research area, many efforts have been made to propose and analyse new prediction methods that could result in better prediction accuracy. Most of the works focused on how to improve the prediction accuracy. There is a lack of coherent and comprehensive research that identifies and analyse the reason why some methods are good predictors when it comes to some of the networks but very inaccurate ones when some other networks are considered.

We address this problem, by exploring the potential correlation between network metrics and prediction accuracy of different methods. We expect

that such approach will enable to find the reasons why methods performance varies on different networks. Apart from having a further understanding of the prediction methods, the study is also important as a theoretical base for developing new prediction method. This could be relevant to many subjects. The prediction methods could help to find the relationships between proteins which might not be easily observed directly due to the interaction complexity. For example, new interactions can be inferred from the existing known interaction networks [38; 39] which shows a much better performance than prediction purely by chance. Online market targeting can benefit even further from the network prediction. For example, Google and Amazon recommend their customers potential goods and services that they might be interested in which is a kind of link prediction problem where the link between customers and products is predicted. Another subject that could benefit from the study is security. Network prediction study could help target criminal networks [97] that have a significant meaning, especially under current global counter-terrorism environment.

This variety and importance of applications of predictive analytics in networks in the context of real-world, large-scale networked systems shows how important it is to understand why the existing methods perform differently on different networks. This study provides insight into why it is a case and offers potential classification of networks depending on their characteristics into those that can be easily predicted and into those where almost all prediction methods fail.

In this chapter, we will introduce the correlation study between prediction accuracy and network metrics as mention in Section 3.3. The experiment result will then be analysed.

5.2 Experiment Design

This experiment is performed on six selected networks. The description of network selection and preparation can be found in Chapter 4.2.1. In order to be able to apply selected methods and taking into account the types of datasets available, the network is represented as a binary un-weighted network. This enables consistent and comprehensive review of the existing metrics and prediction methods.

The methodology followed in the experiment is presented in Figure 3.2. The experiment steps includes:

1. Calculate the selected network metrics for the processed training dataset of each selected networks.

2. Predict links using the selected prediction methods formed from training dataset to test dataset.

3. For each network metrics, calculate its correlation coefficient with all the link prediction results.

4. Analysis the correlation coefficient results.

For the training set of each selected network, the network metrics are calculated with toolboxes provided by KONECT [88]. The selected prediction methods will be applied to each of the processed training set and the accuracy of each method on each dataset is measured using AUC. For the implementation of those methods, we applied the toolbox that presented in [31] and all the experiments were implemented in Matlab.

Once the data of prediction accuracy for each method and the metrics of each network are calculated, the correlation between them will be analysed by calculating the Pearson's linear correlation coefficient [98].

5.2.1 Network Metrics

The main goal of this study is to explore whether the correlation between the prediction accuracy and network metrics exists. The metrics that are calculated include:

Global Clustering Coefficient (GCC),[42]

It is defined in as:

$$GCC = \frac{3 * \text{number of triangles in the network}}{\text{Number of connected triples of vertices}} \tag{5.1}$$

It shows the transitivity of the network as a whole. The coefficient range is between 0 and 1.

Average Clustering Coefficient (ACC) [6]

It is based on local clustering C_l. For each of the node l, its local clustering coefficient can be calculated by:

$$C_l = \frac{\text{Number of triples connected to node } l}{\text{Number of triples centred on node } l} \tag{5.2}$$

and then the ACC can be calculated as:

$$ACC = \frac{1}{n} \sum_{l} C_l \tag{5.3}$$

where n is the number of nodes in a network.

Network Density (ND) [59]

Network Density is a ratio between existing links and all possible links given

the node numbers.

$$\text{Network Density} = \frac{\text{Number of Existing Links}}{\text{Number of all possible links}} \tag{5.4}$$

where

$$\text{Number of all possible links} = \frac{n * (n-1)}{2} \tag{5.5}$$

where n is the number of nodes in the network.

Gini Coefficient (GC) [99]

In network study Gini Coefficient is defined as:

$$G = \frac{2\sum_{i-1}^{n} ik_i}{n\sum_{i-1}^{n} k_i} - \frac{n+1}{n} \tag{5.6}$$

where $k_1 \leq k_2 \leq \cdots \leq k_n$ is a sorted list of degrees in a network and n is a number of nodes in a network. Its value is between 0 and 1, where 0 denotes all the nodes have the same degree number and 1 denotes dominance of single node.

Diameter [59]

The longest of the shortest paths in the network.

$$Diameter = max_{i,j} d(i,j) \tag{5.7}$$

Where $d(i,j)$ is the shortest path between node i and j.

Average Shortest Path (ASP) [59]

The average number of the shortest paths between each pair of vertices.

$$ASP = \frac{1}{n \cdot (n-1)} \cdot \sum_{i!=j} d(i,j) \tag{5.8}$$

Degree Distribution (DD) [59]

The distribution of all the nodes' degree which is used to determine the type of a network (regular, random, small–world, scale–free, etc.).

$$P_k = \text{proportion of vertices with degree k} \tag{5.9}$$

5.2.2 Prediction Methods

In this experiment, we select ten commonly used prediction methods for the study (Detailed introduction can be found in Section 2.4):

- Common Neighbours (CN),
- Jaccard's Coefficient Index (JI),
- Preferential Attachment (PA),
- Adamic/Adar Index (AA),
- Katz method (Katz),
- Cosine Similarity (Cos),
- Sørensen Index (Sor),
- Hub Promoted Index (HPI),
- Hub Depressed Index (HDI) and
- Leicht-Holme-Newman Index (LHN).

5.3 Analysis of the Relationship Between Network Metrics and Prediction Accuracy of Different Methods

5.3.1 Networks Metrics

The values of network metrics for each of the extracted social network are presented in Table 5.1. As it is much easier to set up relationship between people in online social network than in real world network, the average shortest path in our experiments are all smaller than six, the number suggested by the six degrees of separation theory [100]. The average shortest path of the six selected networks is 3.65. The longest ASP that equals 5.72 is for Facebook network and the shortest ASP is 2.34 for the Flickr network. This reflects the small–world property of the networks. People are closer to each other in online social networks than in face–to–face networks. This phenomenon was also pointed out in [101] where authors established that the average shortest path of Twitter is 3.43.

The degree distributions of the six networks, shown in Fig 5.2, indicates that they are scale-free networks as the distributions follow the power law.

We also compared the GCC and ASP metrics of the real network with the theoretical metrics of random network and regular network that have same number of nodes and links. The analytical formulas for GCC and ASP in random and regular networks with a given number of nodes and links are given in Table 5.2. The results of calculations for each analysed network are presented in Table 5.3.

Fig 5.1 plots the metrics of six analysed networks and related theoretical networks respectively. It shows that the clustering coefficient of the anal-

Datasets	GCC	ACC	Density	Gini	Diameter	ASP
Facebook	0.0341	0.1176	0.0008674	0.473	16	5.7235
Flickr	0.0658	0.3294	0.0219	0.5931	6	2.3447
UC Irvine	0.0197	0.1075	0.0084	0.6394	7	3.0463
PWr	0.0048	0.2666	0.00076	0.6407	16	4.0162
Enron	0.029	0.1946	0.0018	0.7172	10	3.6818
YouTube	0.0286	0.2838	0.003	0.7222	5	3.0709

Table 5.1: Network Metrics Results

	Random Network	Regular Network
GCC	$\frac{k}{n}$	$\frac{3(k-2)}{4(k-1)}$
ASP	$\frac{\log n}{\log k}$	$\frac{n}{2k}$

Table 5.2: Analytical formulas for GCC & ASP in random and regular networks. Note: k is the average degree and n is the number of nodes in the network

ysed networks are all between random and regular networks. Meanwhile, the average shortest path of real-world networks is all very close to the random networks. These two phenomena indicate the small–world property of analysed structures. Taking into account both metrics and node degree distribution, it can be concluded that those networks are a combination of small–world and scale–free networks.

5.3.2 Prediction Results

The prediction results are summarised in Table 5.4. Katz method achieved the best average performance and the overall performance is ranked as: Katz > Preferential Attachment > Adamic-Adar > Common Neighbours > Cosine Similarity > Jaccard Index > Hub Depressed Index > Hub Promoted Index

	Random Network	YouTube	Regular Network
Nodes	6,000	6,000	6,000
Links	54,596	54,596	54,596
GCC	0.0030	0.0286	0.7064
ASP	2.9983	3.0709	164.8500
		UC Irvine	
Nodes	1,666	1,666	1,666
Links	11,582	11,582	11,582
GCC	0.00835	0.0197	0.6919
ASP	2.8186	3.0463	59.9108
		PWr	
Nodes	6,335	6,335	6,335
Links	15,334	15,334	15,334
GCC	0.0008	0.0048	0.5547
ASP	5.5499	4.0162	654.3060
		Flickr	
Nodes	5,949	5,949	5,949
Links	387,719	387,719	387,719
GCC	0.0219	0.0658	0.7442
ASP	1.7845	2.3447	22.8198
		Facebook	
Nodes	5,784	5,784	5,784
Links	14,507	14,507	14,507
GCC	0.0009	0.0341	0.5633
ASP	5.3717	5.7235	576.5205
		Enron	
Nodes	5208	5208	5208
Links	23977	23977	23977
GCC	0.0018	0.0290	0.6586
ASP	3.8548	3.6818	282.8037

Table 5.3: Theoretical GCC & ASP of Random, Real and Regular Network

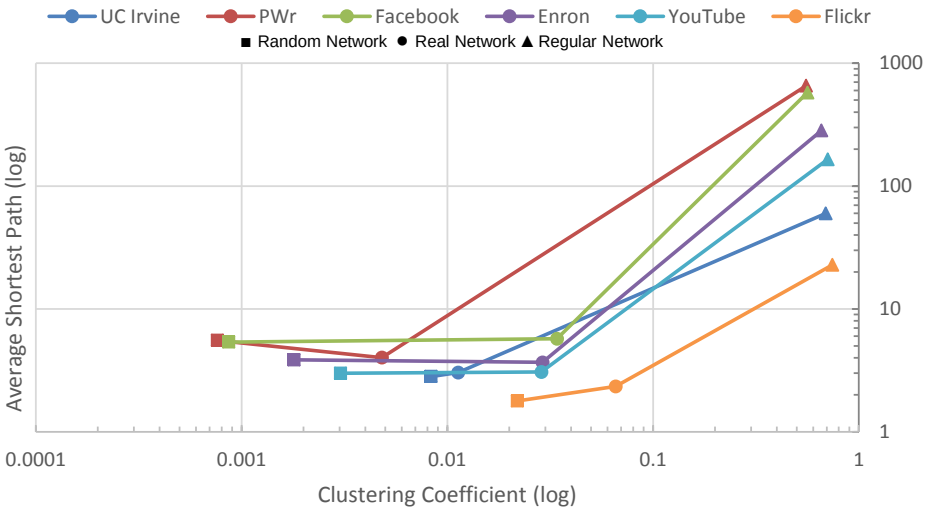
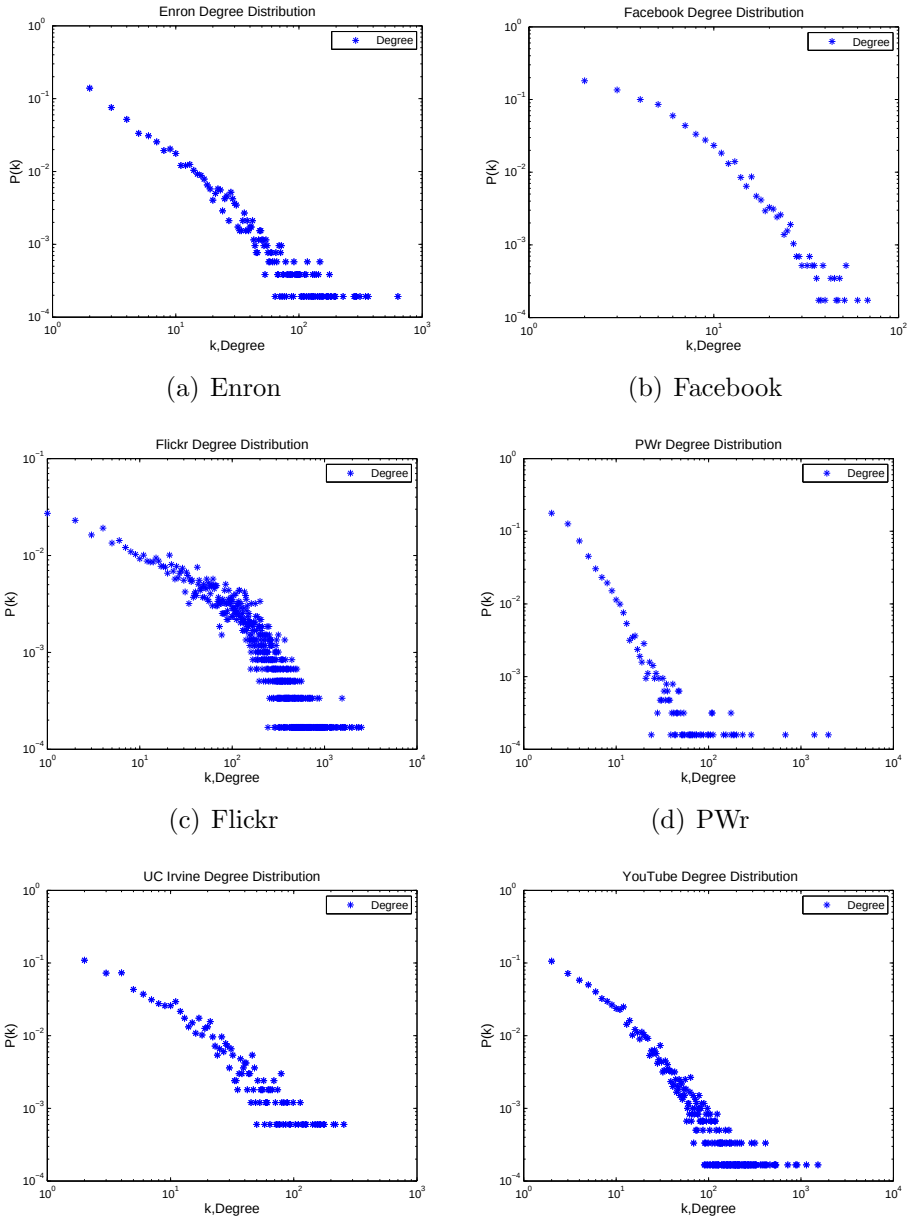


Figure 5.1: Real Network and Theoretical Network Metrics Comparison

> Sørensen > Leicht–Holme–Newman Index. By comparing the variance of each method, we find that the Katz also provides the most stable prediction performance among those methods while Common Neighbours is the worst performing approach. Overall, we find that Katz and Preferential Attachment provide good prediction accuracy together with a relatively stability.



(a) Enron (b) Facebook

(c) Flickr (d) PWr

(e) UC Irvine (f) YouTube

Figure 5.2: Experiment Network Degree Distributions. Note: The degree distributions are all follow the power law with exponent of (a) Enron, $r = 1.85$; (b) Facebook, $r = 1.82$; (c) Flickr, $r = 1.25$; (d) PWr, $r = 2.19$; (e) UC Irvine, $r = 1.56$; (f) YouTube, $r = 1.56$.

Datasets	AUC									
	CN	JI	PA	AA	Katz_β	Cosin	Sor	HPI	HDI	LHN
Facebook	0.67	0.68	0.68	0.68	0.85	0.67	0.67	0.67	0.67	0.67
Flickr	0.89	0.87	0.84	0.89	0.88	0.88	0.87	0.84	0.85	0.69
UC Irvine	0.66	0.64	0.84	0.67	0.80	0.64	0.64	0.63	0.64	0.63
PWr	0.68	0.65	0.79	0.69	0.80	0.65	0.65	0.64	0.65	0.64
Enron	0.82	0.79	0.90	0.82	0.93	0.79	0.80	0.79	0.80	0.79
YouTube	0.85	0.80	0.91	0.86	0.92	0.79	0.75	0.80	0.80	0.76
Average	0.76	0.74	0.83	0.77	0.86	0.74	0.73	0.73	0.74	0.70
Variance	0.0105	0.0091	0.0071	0.0099	0.0032	0.0095	0.0084	0.0087	0.0083	0.0041

Table 5.4: Prediction Methods Accuracy Result (AUC)). Note: we choose $\beta = 0 : 0005$ for Katz_β

To study the prediction results from the perspective of each network please see Fig. 5.3. The prediction results of different methods align on the vertical lines for each network respectively. From this figure, we find that for some networks, most of the prediction methods result in a good prediction accuracy. Such networks include Flickr, Enron and YouTube. We call this type of networks the 'prediction friendly' network. Apart from this type of network, there are also some networks for which most of the prediction approaches provide fairly low accuracy, such as Facebook, UC Irvine and PWr. Similarly, we call those network 'prediction unfriendly' networks. Please note that in the experiments, for both prediction friendly and unfriendly networks, Katz_β always provide a good performance level.

5.3.3 Correlation between Prediction Accuracy and Network Metrics

Table 5.5 shows the Pearson's linear correlation coefficient of prediction accuracy and network metrics. The closer the absolute value to 1, the higher the correlation between analysed factors is. In our experiment, the Preferential Attachment and Gini Coefficient provides the highest correlation coefficient, – 0.94, – which indicates that they generally follow a positive linear relationship. Cosine–GCC and Sor–GCC also provide a correlation coefficient above 0.8. The Diameter and Average Shortest Path shows a negative linear correlation to all the prediction methods meaning that the smaller the diameter and the shorter the ASP, the better prediction results. Fig.5.4 presents a heat–map plot to show the degree of linear relation between the two factors where we use the absolute value of Correlation Coefficient. It should be clear that this correlation coefficient does not indicate the accuracy of the method. For example, although the prediction method Katz does not show strong correlation to any

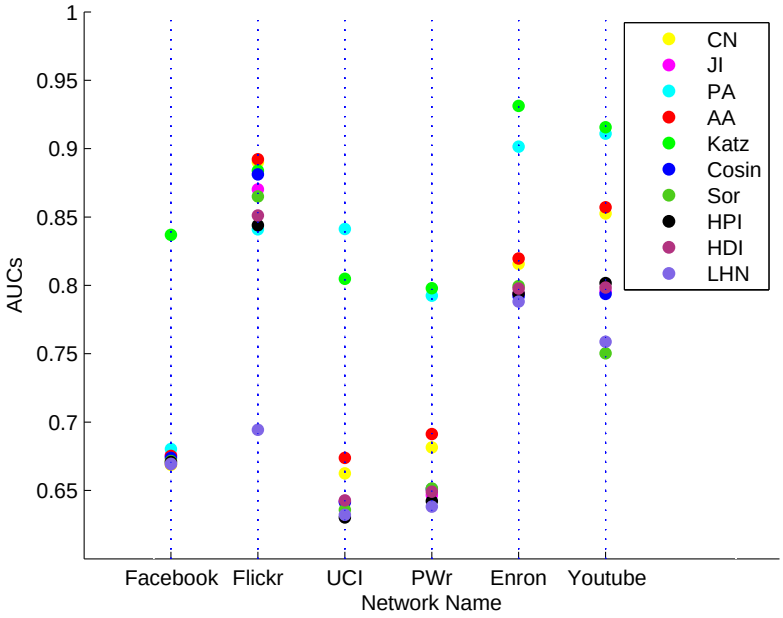


Figure 5.3: The AUC Prediction Results for Each Network

of the network metrics, it still provides best result in our experiments. The results can be found in Table 5.4, where it is shown that Katz always provides a high prediction accuracy regardless the tested network.

The most important value of our correlation study lies in the variety of prediction methods used in the experiments. The prediction with methods combination could be a way to improve accuracy and this is investigated in the Chapter 6.

	CN	JI	PA	AA	Katz$_\beta$	Cosine	Sor	HPI	HDI	LHN	AVG
GCC	0.68	0.79	0.05	0.68	0.47	0.80	0.81	0.73	0.74	0.27	0.60
ACC	0.75	0.68	0.43	0.76	0.39	0.70	0.65	0.67	0.68	0.30	0.60
Density	0.52	0.58	0.18	0.52	0.09	0.61	0.61	0.48	0.52	-0.12	0.40
Gini	0.45	0.30	0.94	0.46	0.49	0.29	0.25	0.36	0.37	0.57	0.45
Diameter	-0.67	-0.61	-0.77	-0.68	-0.51	-0.61	-0.52	-0.61	-0.63	-0.39	-0.60
ASP	-0.63	-0.55	-0.79	-0.65	-0.29	-0.57	-0.52	-0.52	-0.56	-0.18	-0.53

Table 5.5: Pearson Correlation of Prediction Methods Accuracy and Network Metrics
This table shows the correlation between prediction methods accuracy and network metrics calculated with Pearson's linear correlation coefficient. The number within the range of [-1,1] where 1 is completely positive correlation, 0 is no correlation, and -1 is completely negative correlation.

Dataset	GCC	ACC	Diameter	ASP	Ave Rank
PWr	6	3	5	5	4.75
Facebook	2	5	5	6	4.5
UC Irvine	5	6	3	2	4
Enron	3	4	4	4	3.75
YouTube	4	2	1	3	2.5
Flickr	1	1	2	1	1.25

Table 5.6: Metrics Rank of Networks

Table 5.5 also shows the average correlation of network metrics and prediction accuracy. As we know the closer the absolute value of correlation to 1, the stronger the linear relation. Here we take 0.5 as a threshold for strong correlation. According to this, we find that there are four metrics strongly correlated with the prediction accuracy which includes GCC, ACC, Diameter and ASP. So it is reasonable to assume that these metrics could be used to classify the prediction friendly and unfriendly networks. We ranked each of the analysed networks according to the metrics that have strong correlation with prediction accuracy and based on this for each network we calculate the average ranking (Table 5.6). Top three ranked networks (with the small average ranks) are the prediction friendly networks and the other three are prediction unfriendly networks. It can be seen that the prediction friendly networks usually have large global and local clustering coefficient, a short average shortest path as well as small diameter. It suggests that networks with the structural profile similar to small–world network are easier to predict than networks similar to random structures.

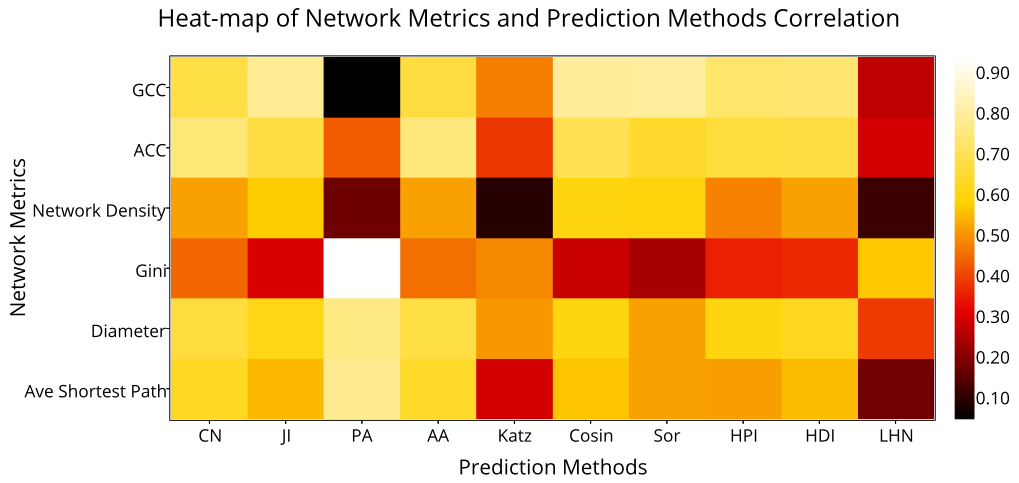


Figure 5.4: Heat-map of Network Metrics and Prediction Methods Correlation. Note: As for the Pearson Coeffcient, both 1 and -1 stands for linear relationship (positive and negative respectively), we use the absolute value of correlation coeffcient in this figure to indicate whether the two factors are linearly correlated.

5.4 Conclusion

In this study, we looked into the correlation between ten prediction methods and different network metrics in six timestamped social networks. The study of network metrics confirmed that the node degree distribution of real world social networks follows a power law distribution. We also found that the average shortest path of online social network is much smaller than six. This might be due to the fact that online relationships are much easier to setup. The results of the prediction accuracy show that the best method among the tested ones is Katz_β. It is also the most stable technique from all tested ones. As it always perform very well, its performance is not highly correlated to the network metrics as shown in Figure 5.4. Preferential Attachment is the second best method that also provides a good prediction accuracy. In addition, for some prediction friendly networks, most of prediction methods could provide a good performance while for some others, called in here as prediction

unfriendly networks, most prediction methods are lack of power The Pearson correlation coefficient enabled to investigate the relationship between network metrics and prediction accuracy. The research showed that some methods are highly correlated with certain network metrics (e.g. PAGini, SorGCC and CosineGcc).

There are several further directions of the presented study. In this study, we adopt Pearson Correlation Coefficient to investigate the relationship between network metrics and prediction accuracy. For future works, we could include other tests like Spearman-Rho Coefficient. Meanwhile, more testing, such as statistical significant and permutation test, could be introduced to make the result more robust. As discovered, for some networks, most prediction methods could provide a good performance which we name them as prediction friendly networks. Similarly, we also find the existence of prediction unfriendly networks. We explored the prediction friendly and unfriendly network classification according to the metrics ranking. The problem is that it does not provide an exact threshold that could be used to classify networks. It is out of scope of this research but is a very interesting topic for another study that we plan to conduct. Based on the results of correlation between network metrics and the prediction accuracy, another possible work is to develop a new prediction method which combines several, existing methods and this approach is further investigated in the next Chapter 6, where a hybrid prediction model is proposed and analysed.

Chapter 6

Hybrid Model

In this chapter, we introduce the hybrid model that proposed by us. The model testing results are then summarized and analyzed. The last section states the results and findings.

6.1 Study Background and Motivation

Many studies have shown that the performance of link prediction methods vary on different networks in different scenarios. For instance, in [20], authors found that the Katz and Preferential Attachment method works well in their experiment on a book sales recommendation network. Authors in [21] claimed that Adamic/Adar method provides the best prediction accuracy on Wikipedia Collaboration Network. The problem is that the performance of methods relies much on the networks topology and that has also been pointed out in [102]. There is no prediction method that works for all networks. A prediction method that could self-adapt to different networks is thus required.

Many of the existing prediction methods work better if the network is growing following the same mechanism over time. For example, the common neighbour approach assumes that links are more likely to appear between nodes

with more common neighbours. Only if the network evolves following this rule the common neighbours' prediction method will give better prediction accuracy than other methods. This applies to other prediction methods as well, e.g. preferential attachment approach. However, a real world network might not evolve following only one rule; it could be the combination of two or more rules.

Starting from this, in this chapter, we propose hybrid model with the assumption that networks are evolving following certain rule or the combination of several rules. By finding this set of rules, we can improve the prediction accuracy.

Data used in this model are time-stamped so we solve time-series prediction problem. We apply two approaches: (i) sliding and (ii) growing window when splitting the data for analysis. The proposed hybrid model combines eight widely used topology based link prediction methods with the assumption that networks evolve following certain mechanisms (we call them rules). Our model predicts links based on the rules that we learnt from the past data about the network. The model has been tested with four real world social networks, Facebook friendship network, Flickr Friendship network, Twitter and Wroclaw University of Technology email communication network. The results show that the hybrid model performs better than the other eight methods applied separately. Our experimental results also show that the two analysed networks are evolving in different ways. So apart from better prediction accuracy, our model can be also used as an approach to analyse dynamics of network evolution.

The rest of this Chapter is organised as follows: in Section 6.2, we introduce the hybrid model. After that, the paper presents methods that were combined in the hybrid model. Section 3.4.1 describes the design of the experiments. We then discuss the results of the experiments in Section 6.4. The last section

concludes the findings.

6.2 Hybrid Link Prediction Model

Much effort has been made to develop new link prediction methods and many of those methods have been proved to perform well on different networks in different scenarios. There is no prediction method that works for all networks. Many of the existing prediction methods work better if the network is growing following the same mechanism over time. For example, the common neighbour approach assumes that links are more likely to appear between nodes with more common neighbours. Only if the network evolves following this rule the common neighbours' prediction method will give better prediction accuracy than other methods. This applies to other prediction methods as well, e.g. preferential attachment approach. However, a real world network might not evolve following only one rule; it could be the combination of two or more rules. Starting from this, we proposed our hybrid model with the assumption that networks are evolving following certain rule or the combination of several rules. By finding the rules, we can improve the prediction accuracy.

Classic topology based link prediction methods work by calculating similarity between nodes[29; 31]. The way how the similarity is calculated varies for different prediction methods. For the prediction purposes dataset is split into two sets, the training set and the test set, where the training set is used to calculate the similarity score for prediction and the result will be verified using the test set. Our approach differs as we consider link prediction as a time-series problem. As shown in Fig.6.1, networks is partitioned into small windows (windows can overlap). We assume the network evolution rule from $Win1$ to $Win2$ remains the same as it is from $Win2$ to $Win3$. Thus, once we learnt the evolution rules from $Win1$ to $Win2$, we can deliver a better predic-

tion performance. Our model is able to work with two scenarios, the growing window and the sliding window. Fig.6.1 shows the growing window scenario in which the next action is to grow the window by one step so that we learn the rules from the change from $Win1 \cup Win2$ to $Win3$ and then use it to predict new links in $Win4$. In the sliding window scenario, the model won't memorize the window but only sliding forward. That is, in the next action, we learn the rules from $Win2$ to $Win3$ and use it to predict new links in $Win4$. For $Win5$ we will repeat the process and learn a new rule from $Win3$ to $Win4$. In this way we enable the method to adapt to the rules that may change over time.

To learn the rules, we need to solve the following optimization problem:

$$min(NL - \sum_{i=1}^{m} w_i S_i) \tag{6.1}$$

Subject to:

$$\sum_{i=1}^{m} w_i = 1$$

$$\forall i \in [1, m] : w_i > 0$$

where NL stands for the new links formed in the window $Win2$ against $Win1$, w_i is the weight assigned to each method, S_i is the similarity score matrix calculated from different selected prediction methods and m is the number of selected prediction methods. In another words, the model linearly combines several prediction methods and the rule is the weight vector for each combined prediction method. In our experiment, we use the Matlab toolbox CVX [103; 104] to solve the optimization problem.

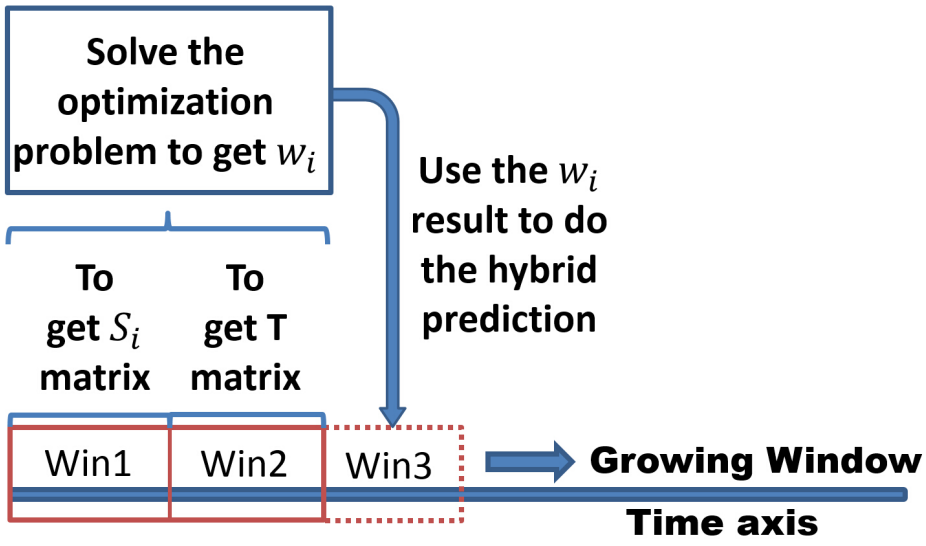


Figure 6.1: Hybrid Prediction Model (Growing Window)

6.2.1 Selected Methods

To test our model, we select eight prediction methods for the study (Detailed introduction can be found in Section 2.4):

- Common Neighbours (CN),
- Jaccard's CoefficientIndex (JI),
- Preferential Attachment (PA),
- Adamic/Adar Index (AA),
- Resource Allocation (RA)
- Cosine Similarity (Cos),
- Sørensen Index (Sor),
- Katz method (Katz).

6.3 Experiment Design

In this study, we followed the methodology introduced in Figure 3.3. The experiments steps for this study is as follows:

1. Find the optimized window size for each networks using the method proposed in [84].
2. Partition each network into the windows and the results are used for the model testing.
3. Solving the optimization problem (as shown in Figure 6.1) to get the weight for each method.
4. Predict links for the next window by linearly combine the selected methods using the weight obtained from the optimization result.
5. Sliding / Growing the windows for next step test. If it is not reach to the end, repeat the process from step 3.

6.3.1 Datasets

To test our hybrid model, we selected four networks. The approach how we select and prepare the networks can be found in Section 4.2.2.

6.3.2 Prediction Accuracy Measures

The prediction performance is measured using precision that stated in Section 3.5.1 and recall, as introduced in Section 3.5.2. Both precision and recall are numbers between 0 and 1. The higher they are, the more accurate the result.

6.4 Experiment Result

Our model and experiments are implemented in Matlab. We run our model for all datasets with both sliding and growing window scenarios. Both the prediction accuracy and weight for each methods are recorded for analysis. To further investigate for what networks the hybrid model is applicable to, we also calculate network topology characteristics.

6.4.1 Prediction Accuracy

Fig 6.2 to Fig 6.23 show the prediction precision and recall results. We calculated the prediction precision and recall for all the scenarios (in Twitter network, we calculated all the daily scenarios). For each dataset, we run our model under different scenarios as indicated in the figure caption. The four sub-charts in each figure depict the prediction precision of eight selected prediction methods as well as our hybrid model. The sub-charts (b), (c) and (d) in each figure depict the prediction precision results when we set N as the number of links we would like to predict. N is an arbitrary number between 0 and average number of new formed links between window steps. The average number of new links is shown in Table 6.1. To make it easier to compare the result between different scenario and networks, we choose N as $100, 500, 1000$ for all datasets apart from Twitter network in the scenario of Monthly Growing and Sliding Windows experiment setting. For Weekly Growing and Sliding Windows experiment, we select N as $50, 100, 500$ for all datasets other than Twitter network. It is because the Twitter network we selected is very small compare to other networks. Thus for Twitter network, we chose N as $10, 50, 100$ for our experiments. The (a) original sub-chart depicts the experiment result if we assume that there is the same number of new links formed next time step as the previous window step. So in the case of sub-chart (a), we predict a

changing number of new links rather than a constant number of links as in the experiment for sub-charts (b), (c) and (d). For the Twitter network, we only run it with the setting of daily sliding and growing windows. The number of links we predicted is $10, 50, 100$ with considering the network size.

Scenarios	Facebook	PWr	Flickr	Twitter
Average Weekly Growing Window New Links	784	732	1119	N/A
Average Weekly Sliding Window New Links	784	1003	1263	N/A
Average Monthly Growing Window New Links	1815	3763	1636	N/A
Average Monthly Sliding Window New Links	1815	4142	1830	N/A
Average Daily Growing Window New Links	N/A	N/A	N/A	245
Average Daily Sliding Window New Links	N/A	N/A	N/A	280

Table 6.1: Average Number of New Links

In all of the sub-charts from Fig 6.2 to Fig 6.23 the prediction precisions of the hybrid model are better or equal to the highest precision and recall results obtained from the eight selected prediction methods separately. Thus we can say that in our experiment, the hybrid model could always deliver the best prediction result. We can also observe that the prediction precision trend of the hybrid model is similar to other methods. That is to say if other methods perform well (or poor) in one window step, our hybrid model performs well (or poor) too. This is not surprise as the hybrid model is a combination of other methods. It cannot predict new links other than the links predicted by combined methods meaning that the hybrid model has its limit.

6.4.2 Facebook Friendship Network

The prediction results for Facebook network is shown in figures Fig 6.2 – Fig 6.7. For the monthly window setting, for both sliding and growing scenario, the hybrid prediction method gives the highest precision result, 0.11 and 0.09 for monthly growing and sliding window respectively when the Top 100 links is predicted. Table 6.2 shows that on average the best precision is for the prediction of Top 100 links - precision of 0.05 for monthly sliding and 0.063 for monthly growing window. Both the highest precision and average precision drop in the scenario of sliding and growing windows as we increase the number of links we are predicting. Our hybrid model performs better when predicting less number of links. The optimal number of links that the hybrid model could predict with the highest prediction accuracy is out of the scope of this study, but it is another interesting topic for future work. For weekly window setting, the highest precision, for both sliding and growing windows, is when Top 50 links is predicted. For the former one it is 0.12 and for the latter one 0.08. We can also see that the standard deviation of the hybrid model prediction result is

also the highest among all the results. It means that the hybrid model results fluctuate heavier than other methods but in the same time they are always above or equal to other results. However, larger standard deviation in this case does not mean our model is not reliable. The last row in Table 6.2 states the improvement rate of our hybrid model over the best performed prediction method (in bold font) among selected 8 methods. We can see that the hybrid prediction model outperforms other methods at least by 54% and in one case the improvement rate is as high as 124%.

We can also observe that, for monthly and weekly window setting, the hybrid model performs better in growing window scenario than in the sliding window one. This is due to the fact that in the growing window scenario, the network topology information is aggregated so that the network information is richer in comparison to that in the sliding window scenario. The richer information helps the model to achieve better prediction result. Similarly, one may think that as window grows, the network topology information gets richer so that the prediction precision should be getting better and better. However, we do not observe a significant increase of precision as window grows for both weekly and monthly experimental settings.

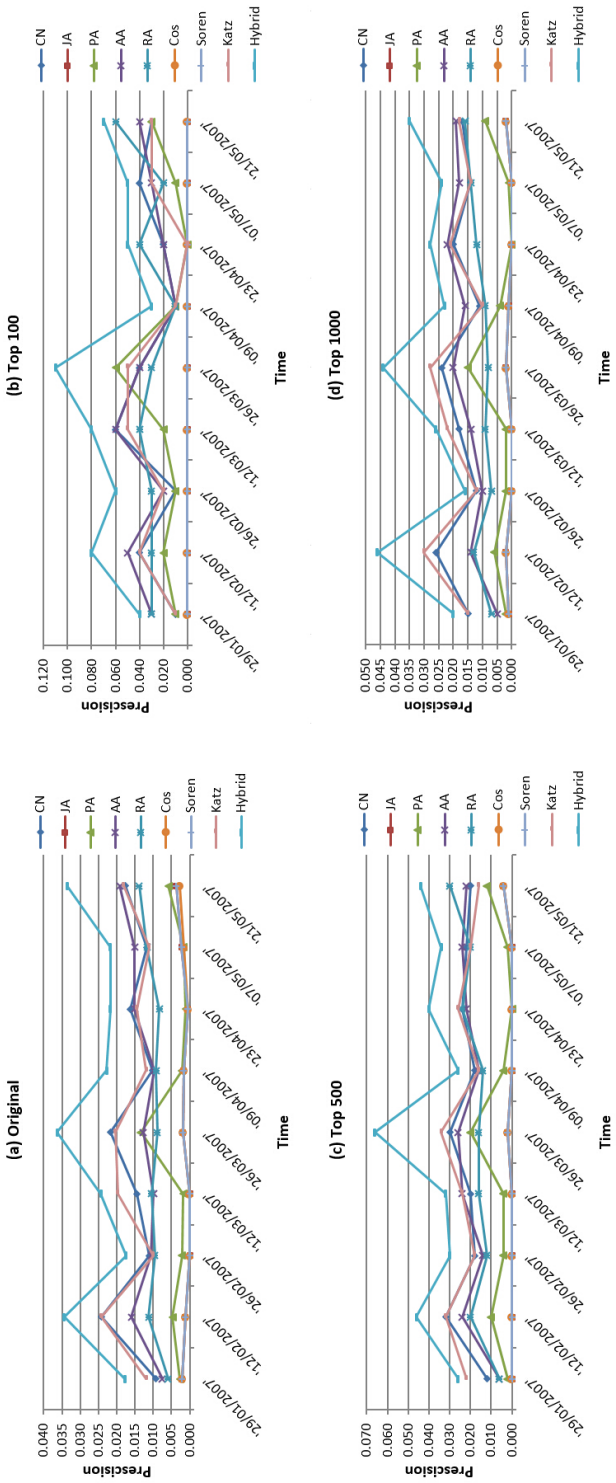


Figure 6.2: Facebook Monthly Growing Window Prediction Precision Result

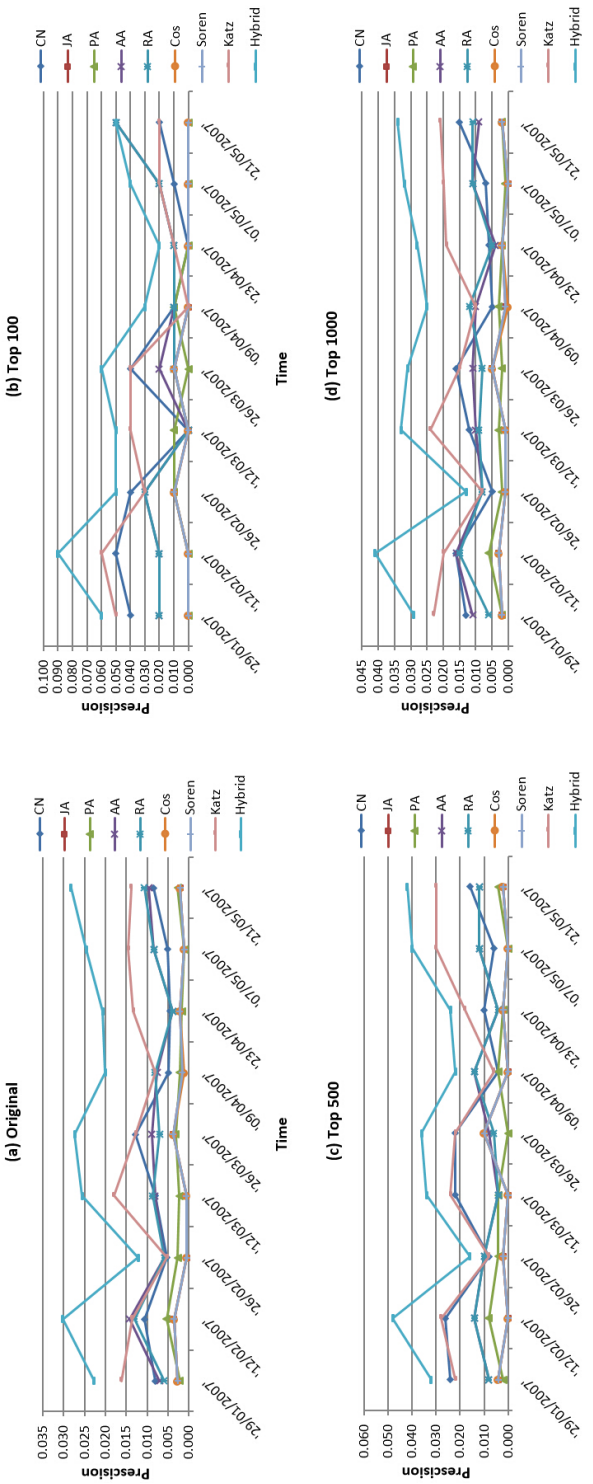


Figure 6.3: Facebook Monthly Sliding Window Prediction Precision Result

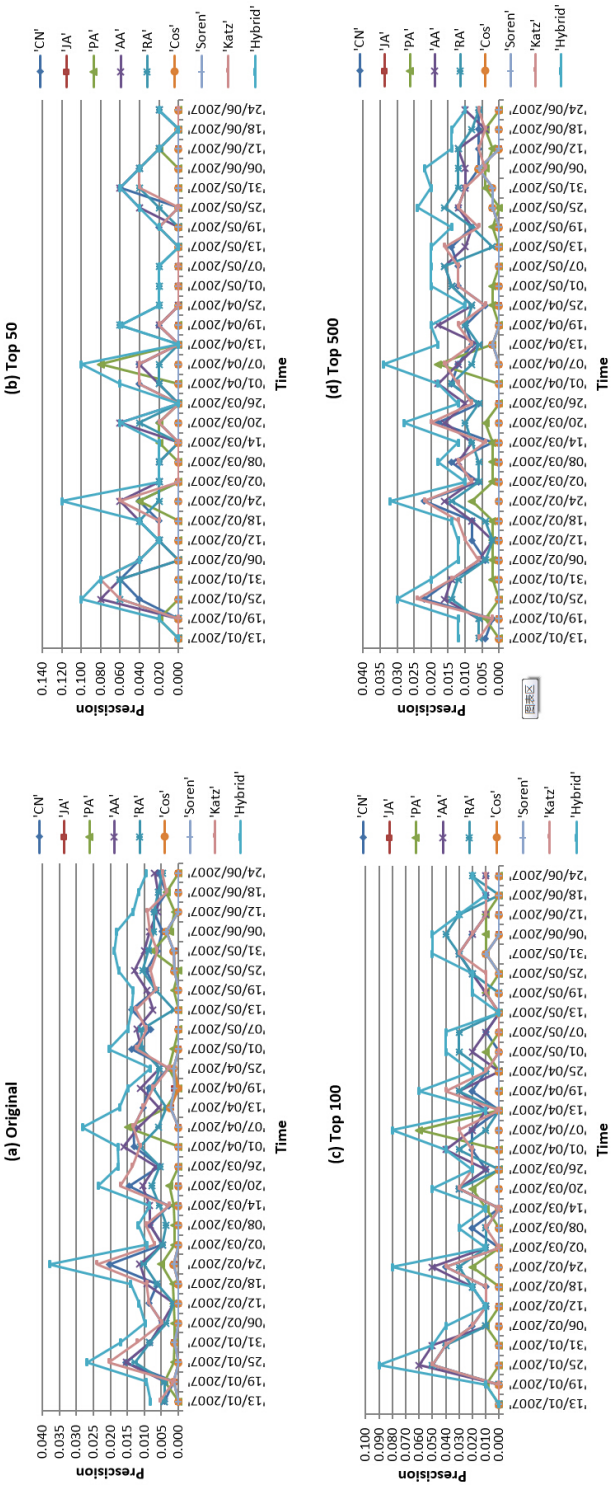


Figure 6.4: Facebook Weekly Growing Window Prediction Precision Result

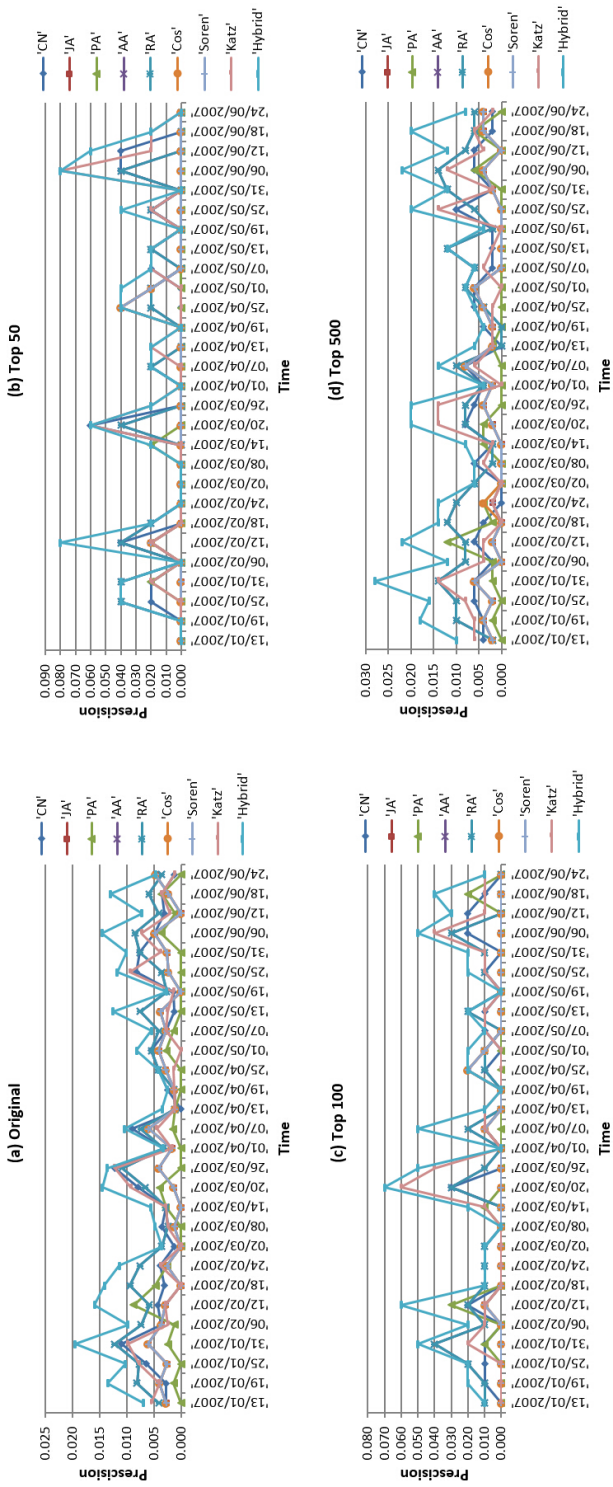


Figure 6.5: Facebook Weekly Sliding Window Prediction Precision Result

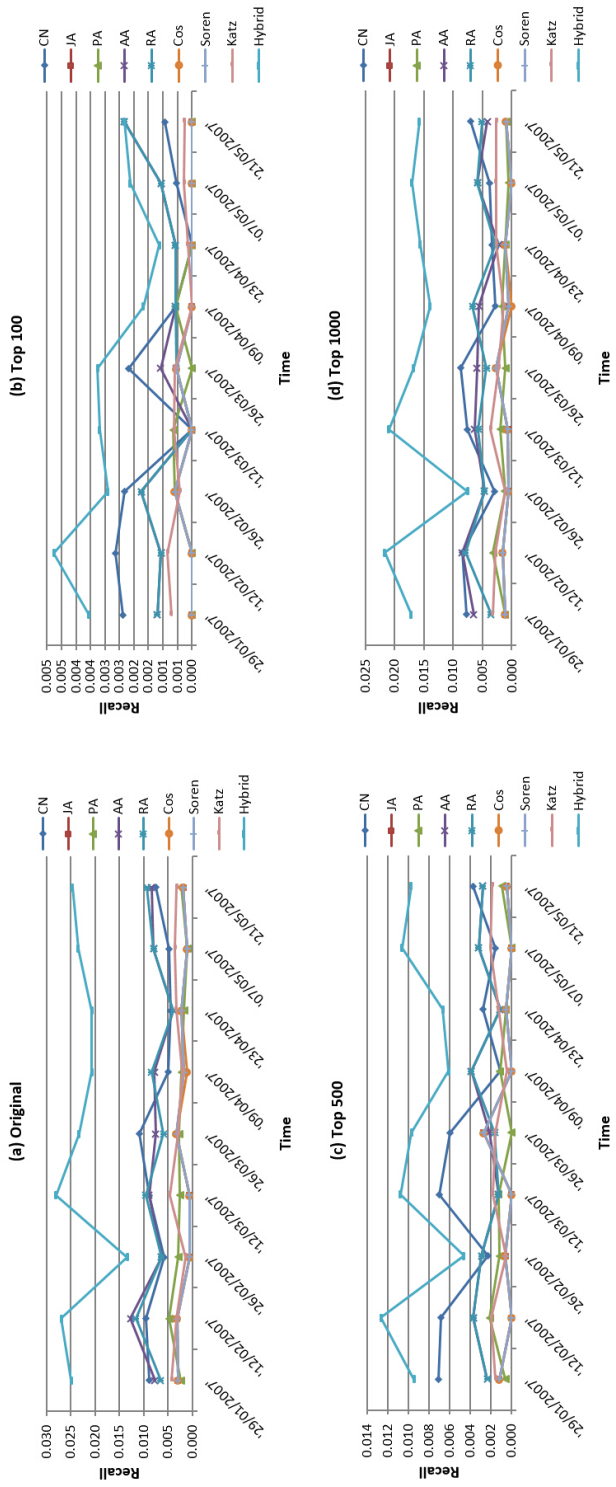


Figure 6.6: Facebook Monthly Sliding Window Prediction Recall Result

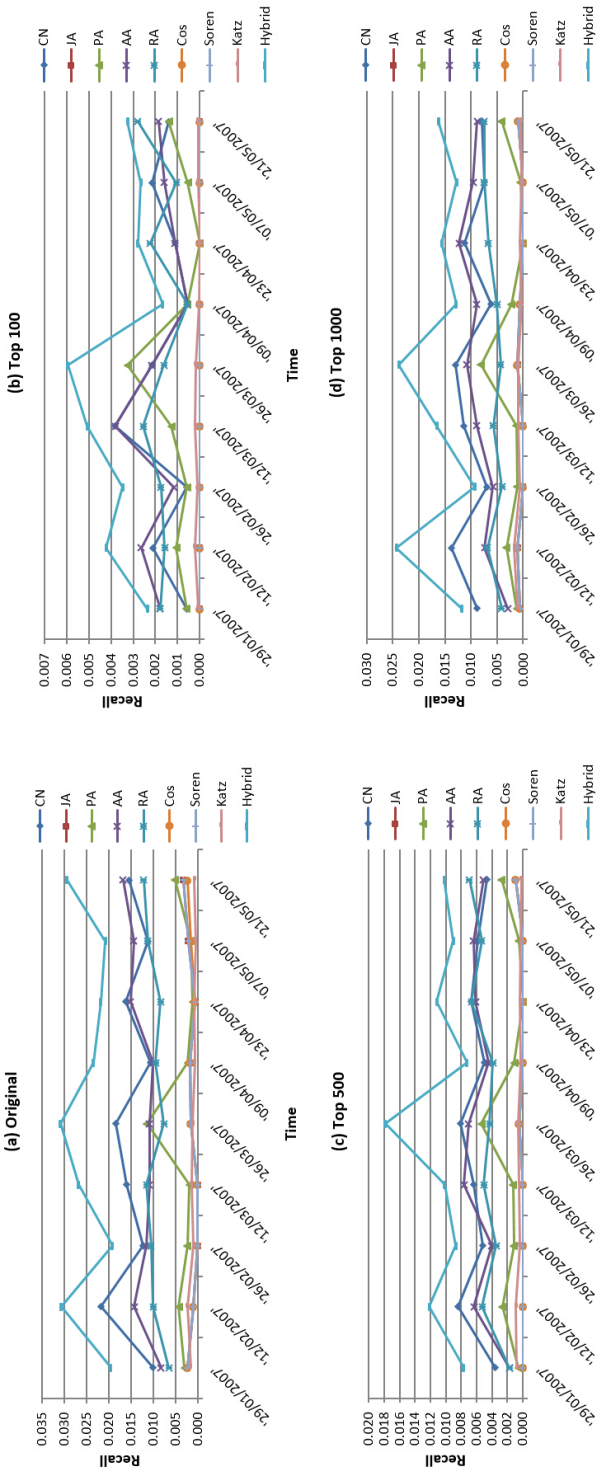


Figure 6.7: Facebook Monthly Growing Window Prediction Recall Result

Table 6.2: Facebook Prediction Average Precision

		Method	Original(std dev)	Top 50(std dev)	Top 100(std dev)	Top 500(std dev)	Top 1000(std dev)
Weekly	Slide	CN	0.0044 (0.0030)	0.0100 (0.0160)	0.0064 (0.0081)	0.0041 (0.0026)	N/A
	Grow		0.0083 (0.0044)	0.0171 (0.0198)	0.0150 (0.0132)	0.0096 (0.0053)	N/A
Monthly	Slide		0.0075 (0.0027)	N/A	0.0233 (0.0183)	0.0153 (0.0080)	0.0106 (0.0045)
	Grow		0.0152 (0.0050)	N/A	0.0289 (0.1500)	0.0218 (0.0060)	0.0174 (0.0050)
Weekly	Slide	JA	0.0025 (0.0017)	0.0029 (0.0088)	0.0018 (0.0047)	0.0022 (0.0022)	N/A
	Grow		0.0005 (0.0009)	0.0000 (0.0000)	0.0004 (0.0019)	0.0004 (0.0012)	N/A
Monthly	Slide		0.0020 (0.0011)	N/A	0.0022 (0.0042)	0.0022 (0.0031)	0.0019 (0.0014)
	Grow		0.0015 (0.0010)	N/A	0.0000 (0.0000)	0.0007 (0.0010)	0.0009 (0.0009)
Weekly	Slide	PA	0.0015 (0.0020)	0.0021 (0.0062)	0.0025 (0.0069)	0.0016 (0.0027)	N/A
	Grow		0.0022 (0.0030)	0.0071 (0.0171)	0.0064 (0.0120)	0.0025 (0.0035)	N/A
Monthly	Slide		0.0026 (0.0011)	N/A	0.0033 (0.0047)	0.0031 (0.0023)	0.0026 (0.0013)
	Grow		0.0039 (0.0036)	N/A	0.0189 (0.0166)	0.0064 (0.0060)	0.0046 (0.0045)
Weekly	Slide	AA	**0.0056** (0.0028)	**0.0114** (0.0155)	**0.0104** (0.0105)	**0.0071** (0.0039)	N/A
	Grow		0.0085 (0.0030)	**0.0221** (0.0240)	**0.0179** (0.0160)	**0.0106** (0.0040)	N/A
Monthly	Slide		0.0082 (0.0027)	N/A	0.0200 (0.0133)	0.0096 (0.0036)	0.0100 (0.0030)
	Grow		0.0129 (0.0040)	N/A	**0.0333** (0.0150)	0.0198 (0.0060)	0.0153 (0.0050)
Weekly	Slide	RA	**0.0056** (0.0028)	**0.0114** (0.0156)	**0.0104** (0.0105)	0.0071 (0.0039)	N/A
	Grow		0.0066 (0.0030)	0.0207 (0.0189)	0.0175 (0.0148)	0.0089 (0.0040)	N/A
Monthly	Slide		0.0080 (0.0026)	N/A	0.0189 (0.0137)	0.0093 (0.0038)	0.0094 (0.0029)
	Grow		0.0099 (0.0020)	N/A	0.0322 (0.0130)	0.0176 (0.0070)	0.0106 (0.0030)
Weekly	Slide	Cos	0.0025 (0.0018)	0.0029 (0.0088)	0.0018 (0.0018)	0.0023 (0.0022)	N/A
	Grow		0.0005 (0.0009)	0.0000 (0.0000)	0.0004 (0.0019)	0.0004 (0.0012)	N/A
Monthly	Slide		0.0020 (0.0012)	N/A	0.0022 (0.0042)	0.0022 (0.0031)	0.0018 (0.0015)
	Grow		0.0013 (0.0009)	N/A	0.0000 (0.0000)	0.0007 (0.0010)	0.0009 (0.0009)
Weekly	Slide	Soren	0.0025 (0.0017)	0.0029 (0.0088)	0.0018 (0.0047)	0.0022 (0.0022)	N/A
	Grow		0.0005 (0.0009)	0.0000 (0.0000)	0.0004 (0.0020)	0.0004 (0.0010)	N/A
Monthly	Slide		0.0020 (0.0011)	N/A	0.0022 (0.0042)	0.0022 (0.003)	0.0019 (0.0014)
	Grow		0.0015 (0.0010)	N/A	0.0000 (0.0000)	0.0007 (0.0010)	0.0009 (0.0009)
Weekly	Slide	Katz	0.0038 (0.0032)	**0.0114** (0.0188)	0.0100 (0.0141)	0.0049 (0.0046)	N/A
	Grow		**0.0094** (0.0051)	0.0186 (0.0226)	0.0154 (0.0145)	0.0103 (0.0055)	N/A
Monthly	Slide		**0.0129** (0.0037)	N/A	**0.0300** (0.0183)	**0.0209** (0.0084)	**0.0178** (0.0053)
	Grow		**0.0158** (0.0050)	N/A	0.0267 (0.0170)	**0.0231** (0.0060)	**0.0189** (0.0070)
Weekly	Slide	Hybrid	*0.0092 (0.0046)*	*0.0229 (0.0243)*	*0.0232 (0.0191)*	*0.0126 (0.0065)*	N/A
	Grow		*0.0158 (0.0068)*	*0.0364 (0.0321)*	*0.0325 (0.0240)*	*0.0179 (0.0067)*	N/A
Monthly	Slide		*0.0235 (0.0051)*	N/A	*0.0500 (0.0189)*	*0.0327 (0.0098)*	*0.0290 (0.0072)*
	Grow		*0.0256 (0.0068)*	N/A	*0.0633 (0.0231)*	*0.0382 (0.0120)*	*0.0291 (0.0098)*
Weekly	Slide	Increase	62%	100%	124%	78%	N/A
	Grow		69%	65%	82%	70%	N/A
Monthly	Slide		83%	N/A	67%	56%	66%
	Grow		62%	N/A	90%	65%	54%

Note: This table summarized the average precision result for Facebook network. Each row represented the average precision for each prediction methods with weekly and monthly prediction in both sliding and grow scenarios with different number we predicted. The numbers in bold are the best performed method amongst select methods in different prediction scenarios.

Table 6.3: Facebook Prediction Average Recall

		Method	Original(std dev)	Top 50(std dev)	Top 100(std dev)	Top 500(std dev)	Top 1000(std dev)
Weekly	Slide	CN	0.0044 (0.0032)	0.0006 (0.0011)	0.0008 (0.0011)	0.0027(0.0017)	N/A
	Grow		0.0082 (0.0043)	0.0011 (0.0012)	0.0019 (0.0016)	0.0062 (0.0034)	N/A
Monthly	Slide		0.0073 (0.0022)	N/A	**0.0013** (0.0010)	**0.0043** (0.0023)	**0.0058** (0.0024)
	Grow		**0.0146** (0.0037)	N/A	0.0016 (0.0010)	**0.0060** (0.0015)	**0.0096** (0.0026)
Weekly	Slide	JA	0.0025 (0.0017)	0.0001 (0.0004)	0.0002 (0.0006)	0.0014 (0.0013)	N/A
	Grow		0.0005 (0.0008)	0.0000 (0.0000)	0.0000 (0.0002)	0.0002 (0.0007)	N/A
Monthly	Slide		0.0019 (0.0010)	N/A	0.0001 (0.0002)	0.0006 (0.0008)	0.0010 (0.0007)
	Grow		0.0014 (0.0010)	N/A	0.0000 (0.0000)	0.0001 (0.0003)	0.0005 (0.0005)
Weekly	Slide	PA	0.0015 (0.0020)	0.0001 (0.0004)	0.0003 (0.0010)	0.0011 (0.0019)	N/A
	Grow		0.0020 (0.0025)	0.0004 (0.0010)	0.0008 (0.0014)	0.0015 (0.0020)	N/A
Monthly	Slide		0.0025 (0.0010)	N/A	0.0002 (0.0003)	0.0009 (0.0006)	0.0014 (0.0007)
	Grow		0.0037 (0.0030)	N/A	0.0010 (0.0009)	0.0017 (0.0016)	0.0024 (0.0024)
Weekly	Slide	AA	**0.0056** (0.0028)	**0.0007** (0.0010)	**0.0013** (0.0013)	**0.0045** (0.0024)	N/A
	Grow		**0.0084** (0.0032)	**0.0014** (0.0015)	**0.0023** (0.0018)	**0.0068** (0.0029)	N/A
Monthly	Slide		**0.0079** (0.0022)	N/A	0.0011 (0.0006)	0.0026 (0.0009)	0.0055 (0.0016)
	Grow		0.0125 (0.0026)	N/A	**0.0019** (0.0009)	0.0055 (0.0017)	0.0084 (0.0026)
Weekly	Slide	RA	**0.0056** (0.0028)	**0.0007** (0.0010)	**0.0013** (0.0012)	**0.0045** (0.0024)	N/A
	Grow		0.0065 (0.0028)	0.0013 (0.0012)	0.0022 (0.0018)	0.0057 (0.0025)	N/A
Monthly	Slide		0.0077 (0.0021)	N/A	0.0010 (0.0007)	0.0025 (0.0010)	0.0052 (0.0015)
	Grow		0.0097 (0.0018)	N/A	0.0018 (0.0007)	0.0048 (0.0015)	0.0058 (0.0013)
Weekly	Slide	Cos	0.0025 (0.0017)	0.0002 (0.0006)	0.0002 (0.0006)	0.0015 (0.0014)	N/A
	Grow		0.0005 (0.0008)	0.0000 (0.0000)	0.0000 (0.0002)	0.0024 (0.0007)	N/A
Monthly	Slide		0.0019 (0.0010)	N/A	0.0001 (0.0002)	0.0006 (0.0008)	0.0010 (0.0008)
	Grow		0.0013 (0.0008)	N/A	0.0000 (0.0000)	0.0002 (0.0003)	0.0005 (0.0005)
Weekly	Slide	Soren	0.0025 (0.0017)	0.0002 (0.0006)	0.0002 (0.0006)	0.0014 (0.0013)	N/A
	Grow		0.0005 (0.0008)	0.0000 (0.0000)	0.0000 (0.0002)	0.0002 (0.0007)	N/A
Monthly	Slide		0.0019 (0.0010)	N/A	0.0001 (0.0002)	0.0006 (0.0008)	0.0010 (0.0007)
	Grow		0.0014 (0.0010)	N/A	0.0000 (0.0000)	0.0002 (0.0003)	0.0005 (0.0005)
Weekly	Slide	Katz	0.0016 (0.0014)	0.0003 (0.0005)	0.0005 (0.0008)	0.0013 (0.0012)	N/A
	Grow		0.0004 (0.0004)	0.0000 (0.0001)	0.0000 (0.0001)	0.0003 (0.0003)	N/A
Monthly	Slide		0.0032 (0.0010)	N/A	0.0004 (0.0002)	0.0015 (0.0006)	0.0025 (0.0008)
	Grow		0.0011 (0.0005)	N/A	0.0001 (0.0000)	0.0005 (0.0002)	0.0008 (0.0004)
Weekly	Slide	Hybrid	*0.0091 (0.0047)*	*0.0015 (0.0016)*	*0.0030 (0.0026)*	*0.0080 (0.0043)*	N/A
	Grow		*0.0155 (0.0061)*	*0.0022 (0.0019)*	*0.0041 (0.0029)*	*0.0114 (0.0039)*	N/A
Monthly	Slide		*0.0129 (0.0041)*	N/A	*0.0028 (0.0010)*	*0.0090 (0.0024)*	*0.0123 (0.0039)*
	Grow		*0.0247 (0.0045)*	N/A	*0.0035 (0.0013)*	*0.0105 (0.0030)*	*0.0160 (0.0049)*
Weekly	Slide	Increase	63%	114%	131%	78%	N/A
	Grow		85%	57%	78%	68%	N/A
Monthly	Slide		63%	N/A	115%	109%	112%
	Grow		69%	N/A	84%	75%	67%

Note: This table summarized the average recall result for Facebook network. Each row represented the average recall for each prediction methods with weekly and monthly prediction in both sliding and grow scenarios with different number we predicted. The numbers in bold are the best performed method amongst select methods in different prediction scenarios.

6.4.3 PWr Email Network

Fig 6.8 to Fig 6.13 shows the prediction results for PWr network. Similarly to the results for the Facebook network, the hybrid model always gives the best prediction outcomes. In the monthly experimental setting, the highest precision is obtained when Top 500 links is predicted for growing window scenario (with precision 0.24 and recall 0.12) and when Top 100 links is predicted for sliding window scenario (with precision 0.29 and recall 0.08). The highest precision for weekly window setting for growing and sliding scenarios is observed for Top 50 and Top 100 cases with precision of 0.16 and 0.60 respectively. In growing window scenario for both weekly and monthly experiment setting, we can see that precision increases at the beginning and then precision drop is noticeable as window grows. This is very different from that of Facebook prediction results in which we do not find obvious growth and decline trend. As shown in Table 6.4, on average, the sliding window results are better than the growing window result. The main reason behind this phenomenon is that if there is no reply for an email then the link might not be valid in the future as the proper relationship has not been formed. So if we simply grow the window, the links formed long time ago that are no longer valid just become links that have negative effect on the prediction results. The accumulation of this unwanted effect makes the prediction result very poor as we can see in Fig 6.8 and Fig 6.10 where all precisions are close to 0 as window grows. In Fig 6.9 (b), we can see the prediction results in first window step is much lower than that in Fig 6.9 (a), (c) and (d). This is due to the different number of new links we are predicting. There are more links with similarity score ranking from top 100 to 1000. It gives us a hint that we can improve link prediction accuracy by predicting within a similarity rank range rather than only focusing on the top ranked links. What is more, compared to Fig 6.8 (b), the precision of first

window step in Fig 6.9 (b) is also very low despite both look at Top 100 links. The difference is caused by sliding and growing windows setting as explained above.

The standard deviation of hybrid model prediction in PWr network is similar to that in the Facebook network experiment. The hybrid model prediction precision is always the best and the standard deviation is larger than other methods as well. The improvement of hybrid model over the best precision result among the 8 selected methods is at least 33% and could be as high as 159%.

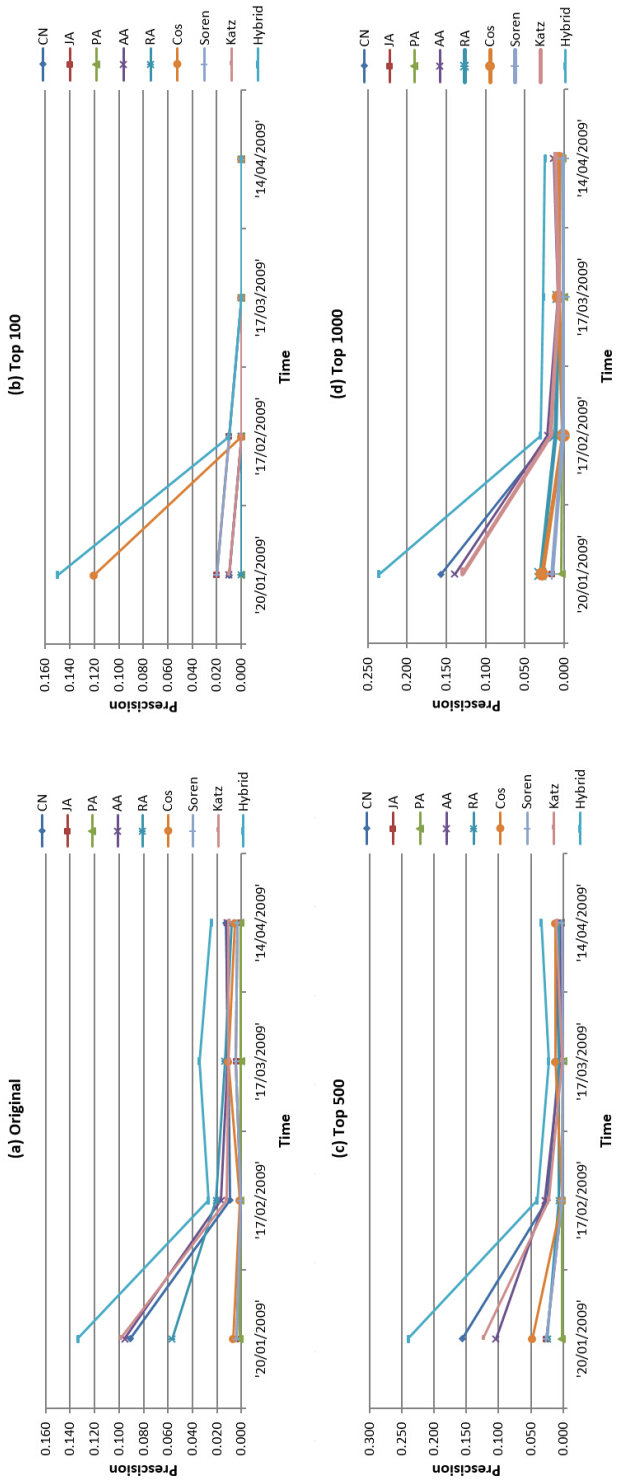


Figure 6.8: PWr Monthly Growing Window Prediction Result

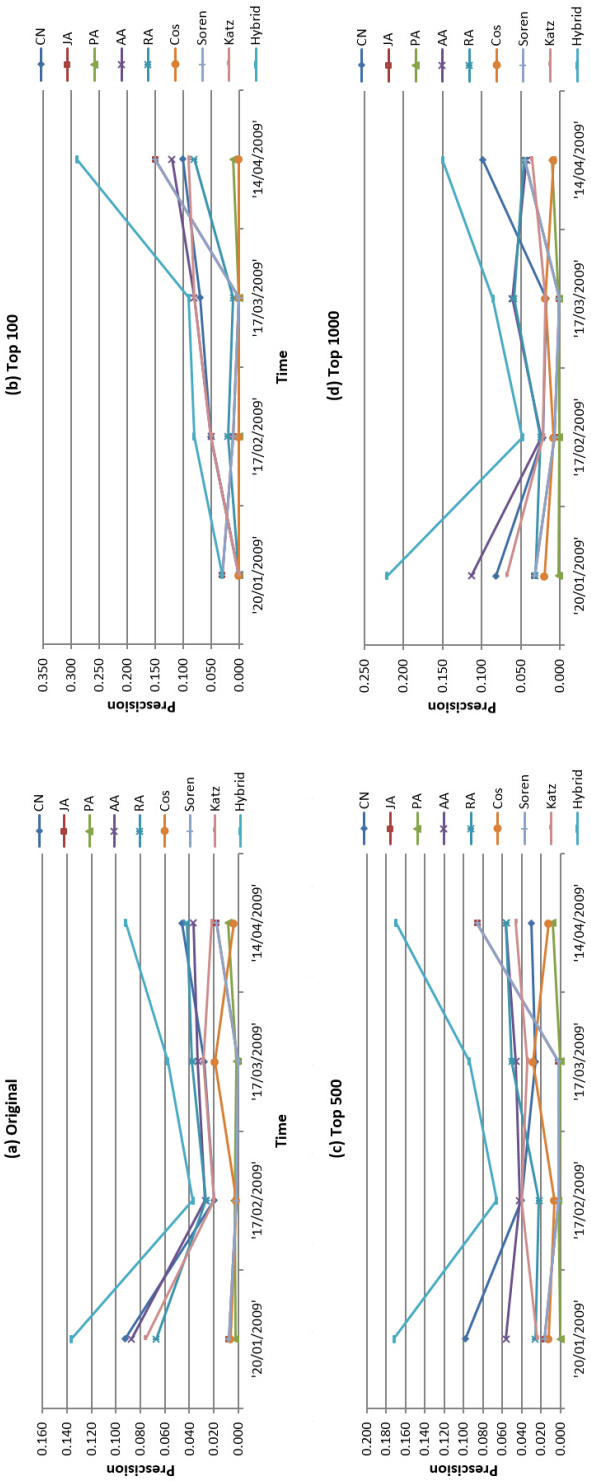


Figure 6.9: PWr Monthly Sliding Window Prediction Result

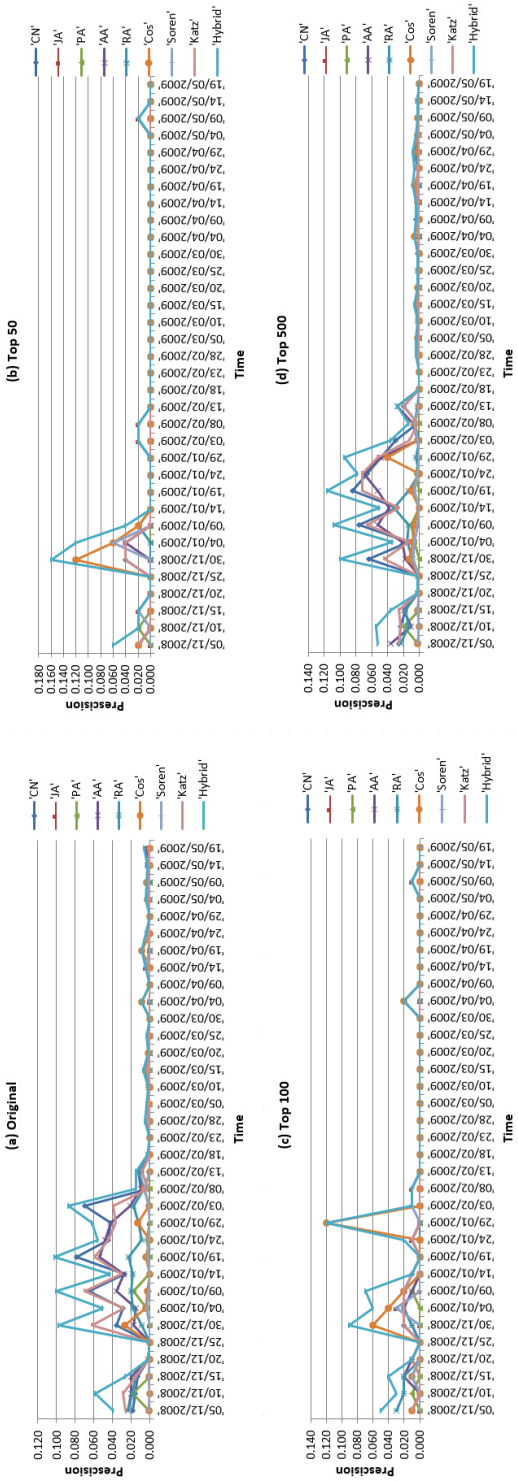


Figure 6.10: PWr Weekly Growing Window Prediction Result

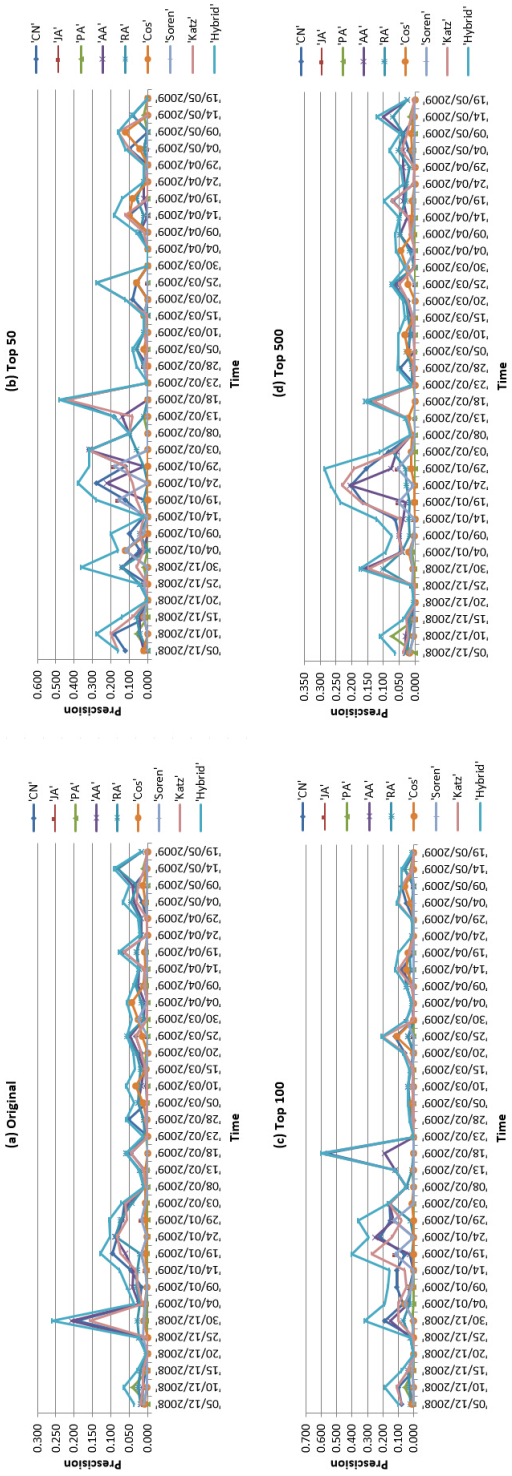


Figure 6.11: PWr Weekly Sliding Window Prediction Result

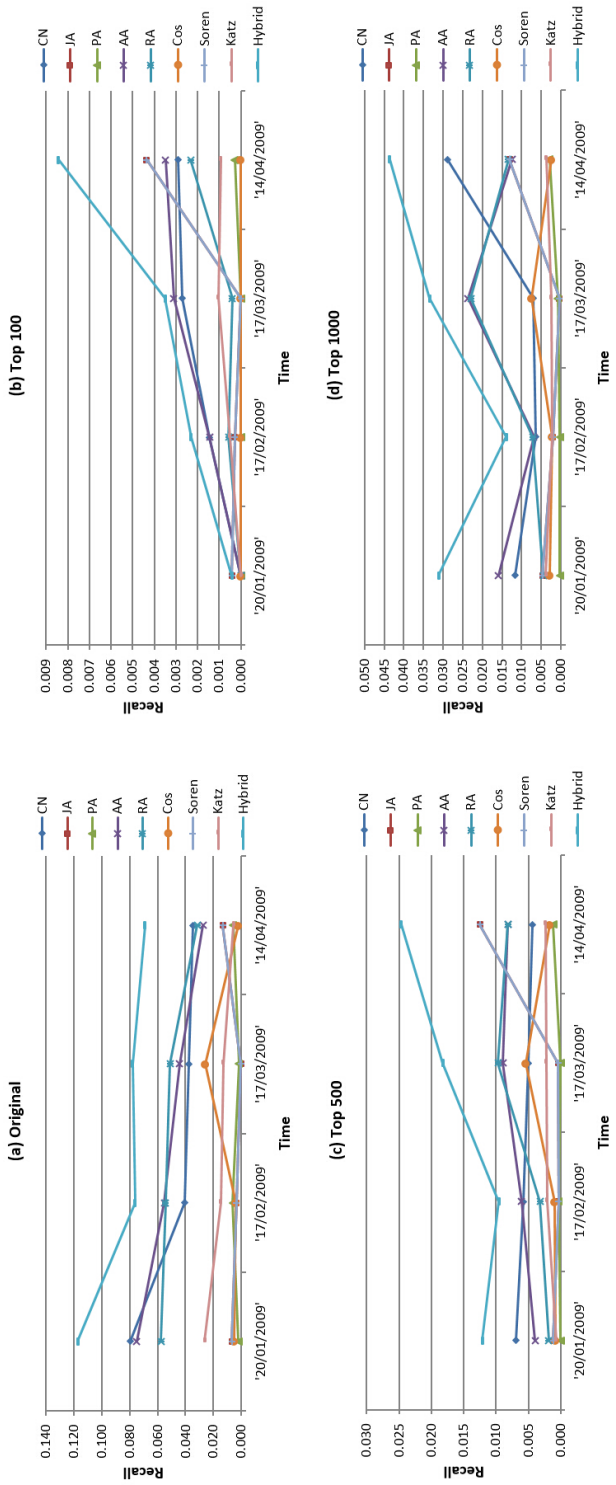


Figure 6.12: PWr Monthly Sliding Window Prediction Recall Result

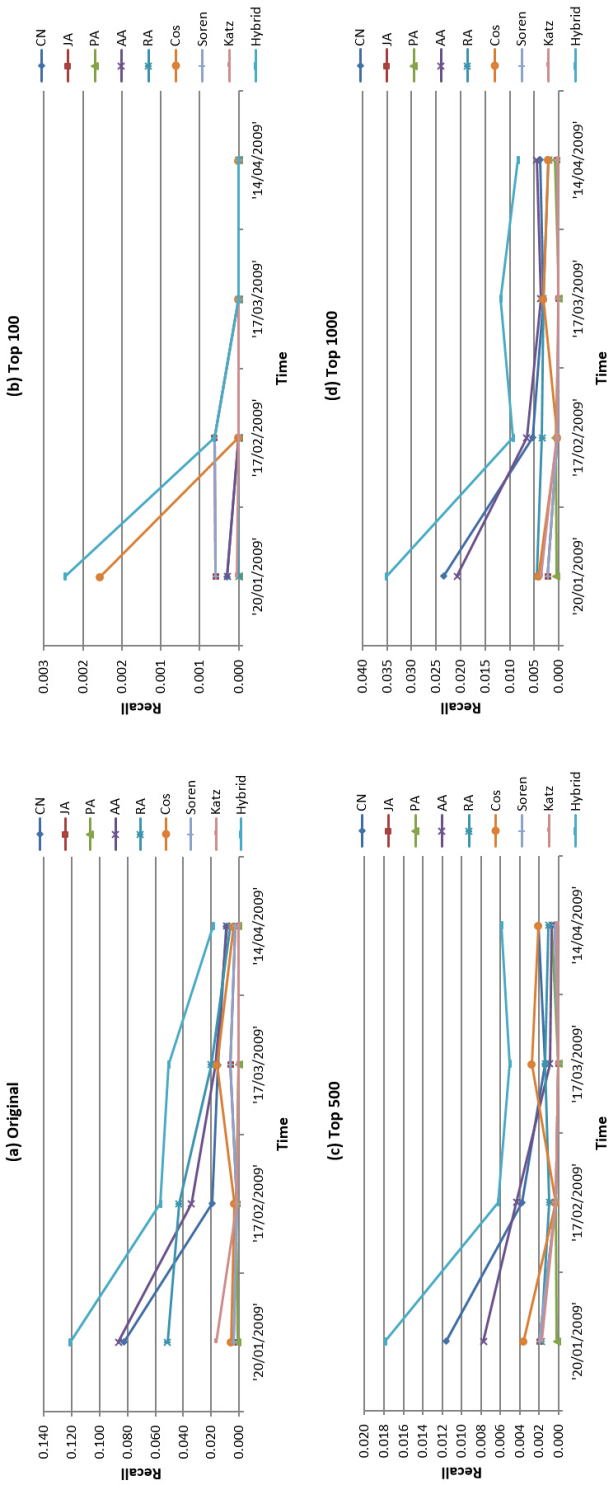


Figure 6.13: PWr Monthly Growing Window Prediction Recall Result

Table 6.4: PWr Prediction Average Precision

		Method	Original(std dev)	Top 50(std dev)	Top 100(std dev)	Top 500(std dev)	Top 1000(std dev)
Weekly	Slide	CN	0.0273 (0.0388)	**0.0735** (0.1043)	0.0703 (0.1069)	0.0423 (0.0547)	N/A
	Grow		0.0136 (0.0219)	0.0029 (0.0099)	0.0029 (0.0062)	**0.0154** (0.0239)	N/A
Monthly	Slide		**0.0466** (0.0281)	N/A	0.0550 (0.0364)	**0.0485** (0.0290)	0.0553 (0.0358)
	Grow		0.0309 (0.0346)	N/A	0.0025 (0.0043)	0.0495 (0.0618)	**0.0480** (0.0630)
Weekly	Slide	JA	0.0040 (0.0059)	0.0165 (0.0438)	0.0124 (0.0311)	0.0068 (0.0131)	N/A
	Grow		0.0007 (0.0014)	0.0041 (0.0117)	0.0024 (0.0060)	0.0012 (0.0021)	N/A
Monthly	Slide		0.0070 (0.0069)	N/A	0.0475 (0.0602)	0.0265 (0.0348)	0.0215 (0.0181)
	Grow		0.0032 (0.0015)	N/A	0.0075 (0.0083)	0.0075 (0.0107)	0.0043 (0.0062)
Weekly	Slide	PA	0.0028 (0.0069)	0.0047 (0.0119)	0.0041 (0.0109)	0.0038 (0.0123)	N/A
	Grow		0.0014 (0.0042)	0.0006 (0.0034)	0.0006 (0.0024)	0.0014 (0.0038)	N/A
Monthly	Slide		0.0038 (0.0026)	N/A	0.0025 (0.0043)	0.0025 (0.0033)	0.0038 (0.0036)
	Grow		0.0010 (0.0004)	N/A	0.0000 (0.0000)	0.0020 (0.0014)	0.0018 (0.0011)
Weekly	Slide	AA	**0.0350** (0.0365)	0.0424 (0.0691)	0.0474 (0.0608)	**0.0439** (0.0465)	N/A
	Grow		0.0107 (0.0151)	0.0018 (0.0075)	0.0024 (0.0055)	0.0124 (0.0191)	N/A
Monthly	Slide		0.0460 (0.0241)	N/A	**0.0625** (0.0438)	0.0500 (0.0062)	**0.0603** (0.0331)
	Grow		**0.0338** (0.0356)	N/A	0.0025 (0.0043)	0.0350 (0.0410)	0.0453 (0.0543)
Weekly	Slide	RA	0.0296 (0.0201)	0.0276 (0.0333)	0.0241 (0.0301)	0.0336 (0.0296)	N/A
	Grow		0.0064 (0.0079)	0.0012 (0.0047)	0.0035 (0.0080)	0.0051 (0.0078)	N/A
Monthly	Slide		0.0434 (0.0148)	N/A	0.0275 (0.0311)	0.0385 (0.0147)	0.0405 (0.0131)
	Grow		0.0248 (0.0190)	N/A	0.0000 (0.0000)	0.0105 (0.0078)	0.0135 (0.0097)
Weekly	Slide	Cos	0.0070 (0.0093)	0.0176 (0.0349)	0.0121 (0.0240)	0.0092 (0.0106)	N/A
	Grow		0.0026 (0.0052)	**0.0065** (0.0226)	**0.0082** (0.0232)	0.0034 (0.0071)	N/A
Monthly	Slide		0.0077 (0.0067)	N/A	0.0000 (0.0000)	0.0145 (0.0082)	0.0138 (0.0058)
	Grow		0.0061 (0.0033)	N/A	**0.0300** (0.0520)	0.0185 (0.0175)	0.0105 (0.0104)
Weekly	Slide	Soren	0.0040 (0.0059)	0.0165 (0.0438)	0.0124 (0.0311)	0.0068 (0.0131)	N/A
	Grow		0.0007 (0.0014)	0.0041 (0.0117)	0.0024 (0.0060)	0.0012 (0.0021)	N/A
Monthly	Slide		0.0070 (0.0069)	N/A	0.0475 (0.0602)	0.0265 (0.0348)	0.0215 (0.0181)
	Grow		0.0032 (0.0015)	N/A	0.0075 (0.0083)	0.0075 (0.0107)	0.0043 (0.0062)
Weekly	Slide	Katz	0.0240 (0.0326)	0.0771 (0.1025)	**0.0756** (0.1118)	0.0435 (0.0594)	N/A
	Grow		**0.0136** (0.0205)	0.0029 (0.0099)	0.0029 (0.0062)	0.0139 (0.0218)	N/A
Monthly	Slide		0.0366 (0.0229)	N/A	0.0550 (0.0350)	0.0360 (0.0081)	0.0363 (0.0194)
	Grow		0.0330 (0.0379)	N/A	0.0025 (0.0043)	**0.0395** (0.0493)	0.0413 (0.0514)
Weekly	Slide	Hybrid	*0.0554 (0.0455)*	*0.1406 (0.1278)*	*0.1256 (0.1326)*	*0.0814 (0.0696)*	N/A
	Grow		*0.0241 (0.0323)*	*0.0141 (0.0345)*	*0.0162 (0.0089)*	*0.0256 (0.0347)*	N/A
Monthly	Slide		*0.0808 (0.0376)*	N/A	*0.1225 (0.0993)*	*0.1255 (0.0466)*	*0.1263 (0.0657)*
	Grow		*0.0549 (0.0456)*	N/A	*0.0400 (0.0636)*	*0.0840 (0.0903)*	*0.0790 (0.0907)*
Weekly	Slide	Increase	58%	91%	66%	85%	N/A
	Grow		77%	118%	97%	67%	N/A
Monthly	Slide		73%	N/A	96%	159%	110%
	Grow		63%	N/A	33%	113%	65%

Note: This table summarized the average precision result for PWr network. Each row represented the average precision for each prediction methods with weekly and monthly prediction in both sliding and grow scenarios with different number we predicted. The numbers in bold are the best performed method amongst select methods in different prediction scenarios.

Table 6.5: PWr Prediction Average Recall

		Method	Original(std dev)	Top 50(std dev)	Top 100(std dev)	Top 500(std dev)	Top 1000(std dev)
Weekly	Slide	CN	0.0273 (0.0395)	**0.0042** (0.0086)	**0.0084** (0.0203)	0.0199(0.0277)	N/A
	Grow		**0.0165** (0.0278)	0.0001 (0.0004)	0.0009 (0.0042)	**0.0082** (0.0102)	N/A
Monthly	Slide		0.0481 (0.0181)	N/A	0.0018 (0.0012)	0.0055 (0.0009)	0.0134 (0.0091)
	Grow		0.0317 (0.0293)	N/A	0.0000 (0.0000)	**0.0047** (0.0041)	**0.0089** (0.0083)
Weekly	Slide	JA	0.0042 (0.0078)	0.0007 (0.0017)	0.0012 (0.0027)	0.0037 (0.0066)	N/A
	Grow		0.0006 (0.0010)	**0.0003** (0.0006)	0.0003 (0.0007)	0.0006 (0.0010)	N/A
Monthly	Slide		0.0061 (0.0048)	N/A	0.0013 (0.0018)	0.0036 (0.0052)	0.0050 (0.0048)
	Grow		0.0036 (0.0018)	N/A	0.0002 (0.0002)	0.0006(0.0008))	0.0007 (0.0009)
Weekly	Slide	PA	0.0023 (0.0051)	0.0002 (0.0006)	0.0004 (0.0010)	0.0017 (0.0050)	N/A
	Grow		0.0013 (0.0032)	0.0000 (0.0002)	0.0000 (0.0002)	0.0007 (0.0017)	N/A
Monthly	Slide		0.0041 (0.0021)	N/A	0.0000 (0.0000)	0.0004 (0.0005)	0.0011 (0.0011)
	Grow		0.0012 (0.0007)	N/A	0.0000 (0.0000)	0.0003 (0.0003)	0.0004 (0.0002)
Weekly	Slide	AA	**0.0380** (0.0382)	0.0023 (0.0042)	0.0052 (0.0077)	**0.0244** (0.0287)	N/A
	Grow		0.0134 (0.0189)	0.0000 (0.0003)	**0.0010** (0.0042)	0.0065 (0.0088)	N/A
Monthly	Slide		**0.0505** (0.0171)	N/A	**0.0020** (0.0014)	**0.0068** (0.0019)	**0.0148** (0.0061)
	Grow		**0.0366** (0.0301)	N/A	0.0000 (0.0000)	0.0034 (0.0028)	0.0088 (0.0069)
Weekly	Slide	RA	0.0374 (0.0381)	0.0016 (0.0018)	0.0030 (0.0042)	0.0230 (0.0270)	N/A
	Grow		0.0088 (0.0113)	0.0000 (0.0002)	0.0004 (0.0010)	0.0036 (0.0056)	N/A
Monthly	Slide		0.0486 (0.0100)	N/A	0.0008 (0.0009)	0.0057 (0.0033)	0.0120 (0.0070)
	Grow		0.0301 (0.0179)	N/A	0.0000 (0.0000)	0.0013 (0.0003)	0.0032 (0.0008)
Weekly	Slide	Cos	0.0082 (0.0133)	0.0011 (0.0022)	0.0015 (0.0034)	0.0068 (0.0115)	N/A
	Grow		0.0020 (0.0042)	**0.0003** (0.0009)	**0.0010** (0.0027)	0.0021 (0.0041)	N/A
Monthly	Slide		0.0095 (0.0094)	N/A	0.0000 (0.0000)	0.0022 (0.0019)	0.0037 (0.0021)
	Grow		0.0073 (0.0050)	N/A	**0.0004** (0.0007)	0.0022 (0.0011)	0.0024 (0.0014)
Weekly	Slide	Soren	0.0042 (0.0078)	0.0007 (0.0017)	0.0012 (0.0027)	0.0037 (0.0066)	N/A
	Grow		0.0006 (0.0011)	**0.0003** (0.0007)	0.0003 (0.0007)	0.0006 (0.0009)	N/A
Monthly	Slide		0.0061 (0.0048)	N/A	0.0013 (0.0018)	0.0036 (0.0052)	0.0050 (0.0049)
	Grow		0.0036 (0.0018)	N/A	0.0002 (0.0002)	0.0006 (0.0008)	0.0007 (0.0008)
Weekly	Slide	Katz	0.0059 (0.0075)	0.0010 (0.0015)	0.0019 (0.0033)	0.0049 (0.0062)	N/A
	Grow		0.0008 (0.0013)	0.0000 (0.0000)	0.0000 (0.0000)	0.0004 (0.0006)	N/A
Monthly	Slide		0.0149 (0.0074)	N/A	0.0006 (0.0004)	0.0018 (0.0007)	0.0031 (0.0007)
	Grow		0.0049 (0.0068)	N/A	0.0000 (0.0000)	0.0005 (0.0006)	0.0011 (0.0015)
Weekly	Slide	Hybrid	*0.0597 (0.0488)*	*0.0078 (0.0095)*	*0.0140 (0.0213)*	*0.0443 (0.0329)*	N/A
	Grow		*0.0269 (0.0365)*	*0.0006 (0.0014)*	*0.0025 (0.0049)*	*0.0141 (0.0144)*	N/A
Monthly	Slide		*0.0849 (0.0188)*	N/A	*0.0036 (0.0029)*	*0.0162 (0.0059)*	*0.0305 (0.0107)*
	Grow		*0.0615 (0.0371)*	N/A	*0.0006 (0.0009)*	*0.0087 (0.0052)*	*0.0161 (0.0110)*
Weekly	Slide	Increase	57%	86%	67%	82%	N/A
	Grow		63%	100%	150%	72%	N/A
Monthly	Slide		68%	N/A	80%	138%	106%
	Grow		68%	N/A	50%	85%	81%

Note: This table summarized the average recall result for PWr network. Each row represented the average recall for each prediction methods with weekly and monthly prediction in both sliding and grow scenarios with different number we predicted. The numbers in bold are the best performed method amongst select methods in different prediction scenarios.

6.4.4 Flickr Network

Fig 6.14 to Fig 6.19 shows the prediction results for Flickr network. In monthly setting, the sliding window scenario generally gives better prediction results than the growing window. Also, the best performed prediction method among the eight selected methods is Preferential Attachment while our hybrid model is slightly better than it. From Table 6.6, we can observe the improvement of the hybrid model against the selected methods. For the top 50 links prediction in weekly growing scenario, the hybrid model provides no improvement. Its prediction precision 0.0013 is same as the result from preferential attachment. Apart from this, the hybrid model is capable to improve the prediction by at least 20%. The highest improvement appears for the monthly sliding scenario when we predict 500 links. Our hybrid model prediction precision is 164% better than the perferential attachment prediction precision which is the best performed methods amongst the eight selected methods.

The method weight as shown in Fig 6.32 to Fig 6.35 also shown that the Preferential Attachment plays an important role in describing Flickr Network evolution. This is caused by the nature of Flickr website. Flickr network grows mainly following the Preferential Attachment rule where 'rich get richer'. The popular accounts will attract more people who follow them. This phenomenon is also visible in weekly experiment setting. Popular user on Flickr could attract more followings but the popularity of individual account might change and shift from one user to another. That is to say in one period, user A is more popular so that A gets more new links, but user B might later be more popular so new links then are formed with B. This is a 'late comer' that becomes popular phenomenon. This periodically links gaining could be the reason why sliding window prediction perform better than growing window prediction in Flickr network.

Table 6.6 and Table 6.7 shows the average precision and recall results. In the best case scenario, our hybrid model improved the precision and recall by 164% and 150% respectively in comparison to the selected best performed method.

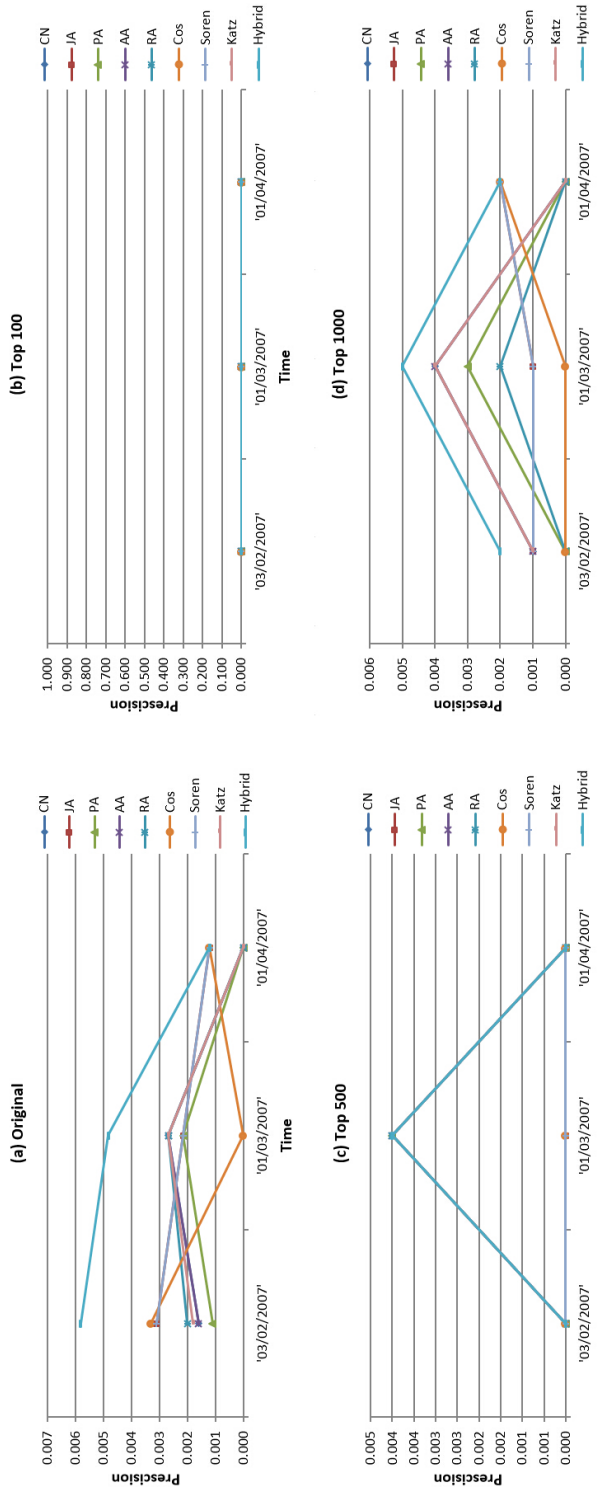


Figure 6.14: Flickr Monthly Growing Window Prediction Precision Result

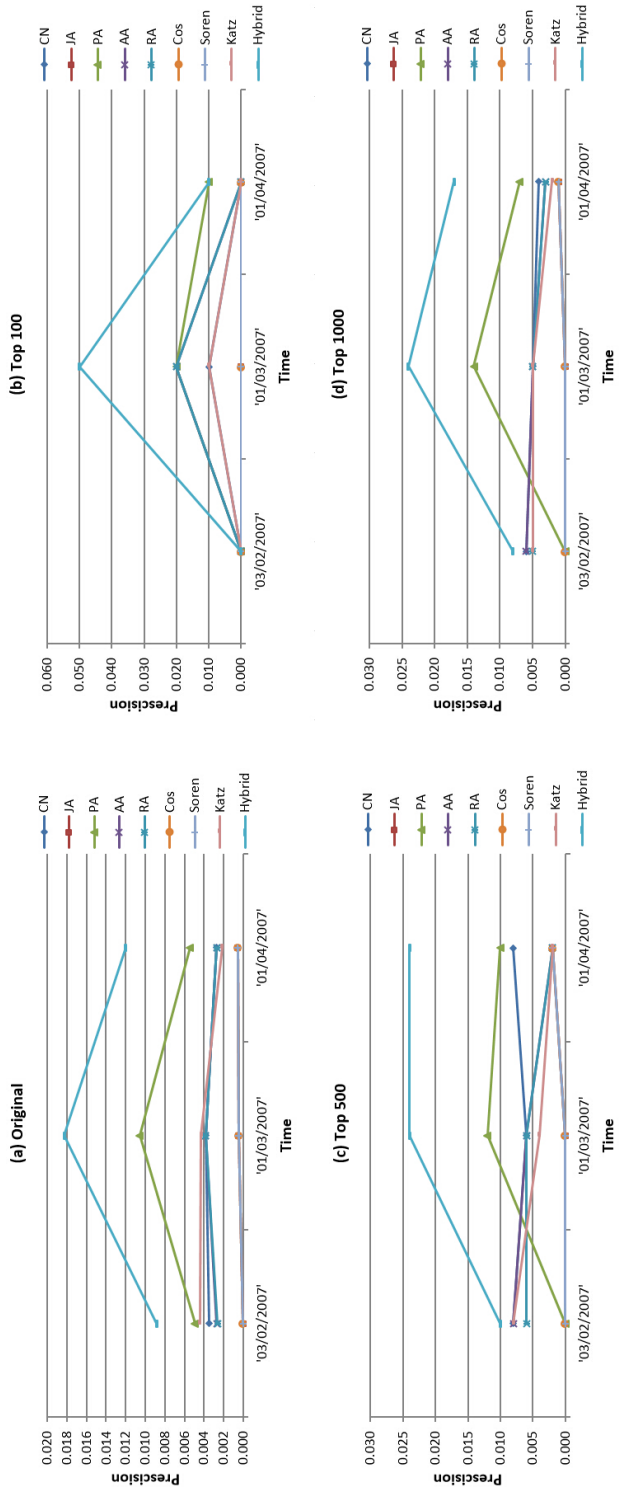


Figure 6.15: Flickr Monthly Sliding Window Prediction Precision Result

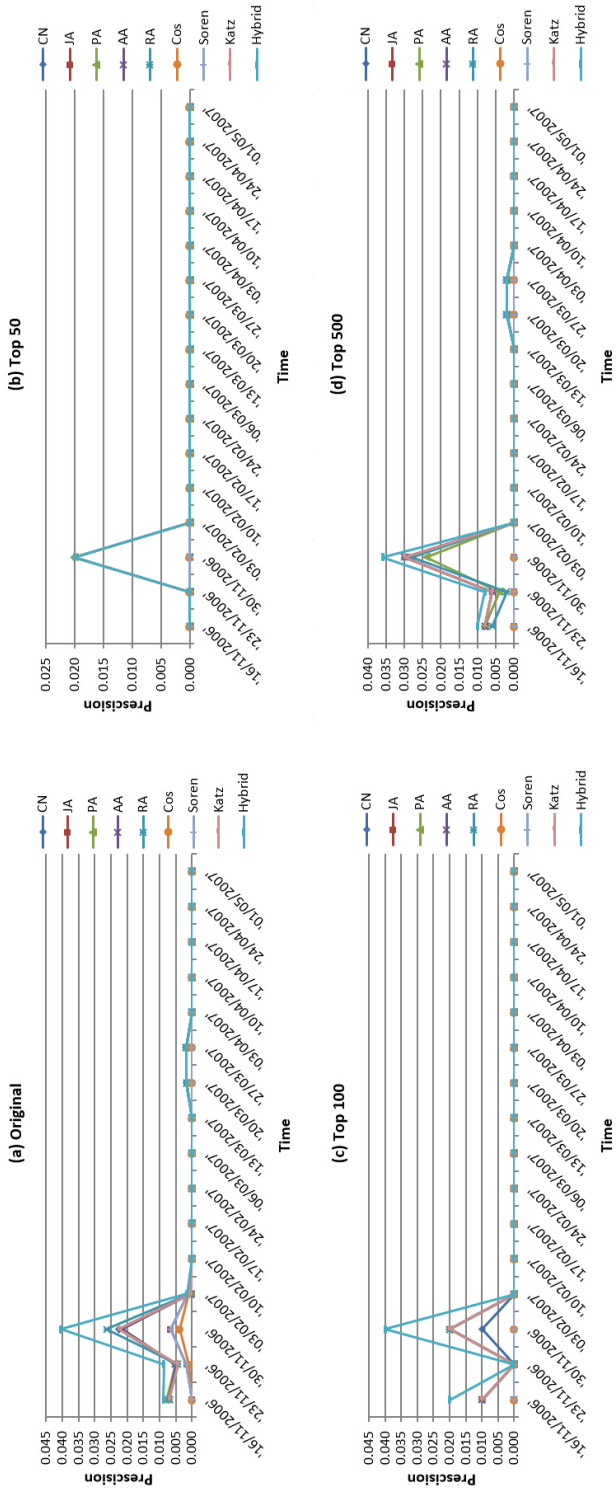


Figure 6.16: Flickr Weekly Growing Window Prediction Precision Result

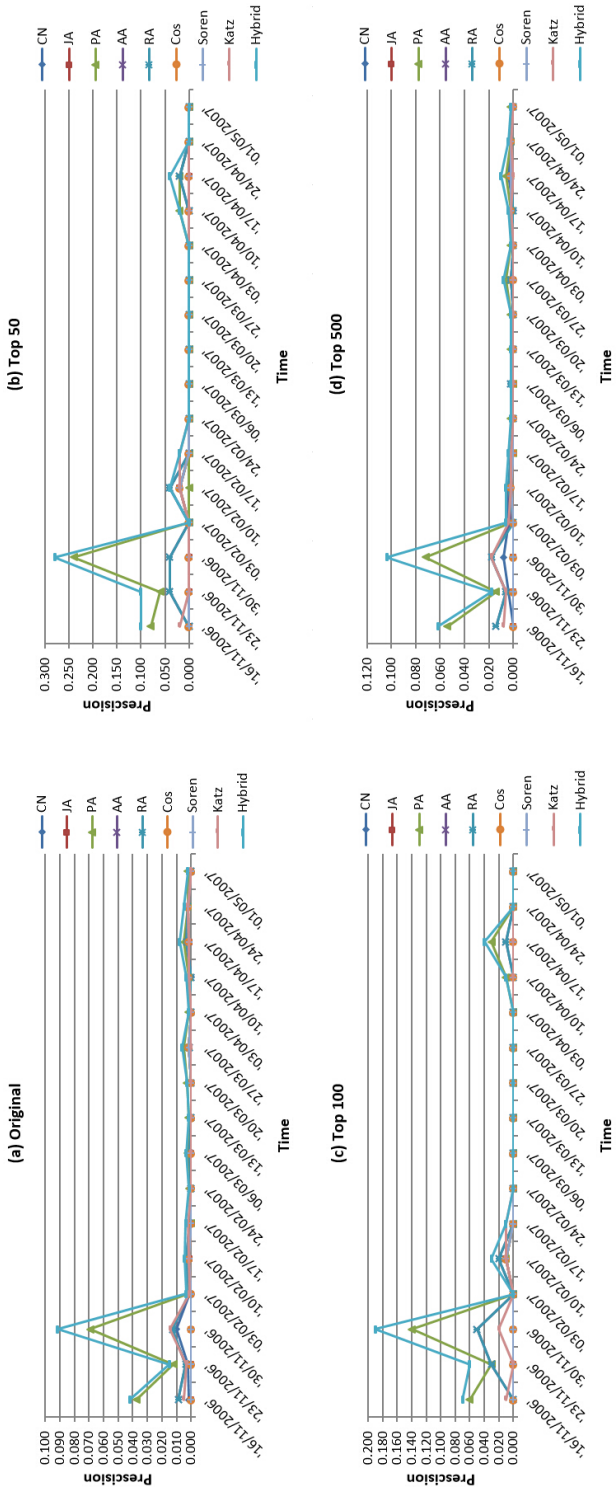


Figure 6.17: Flickr Weekly Sliding Window Prediction Precision Result

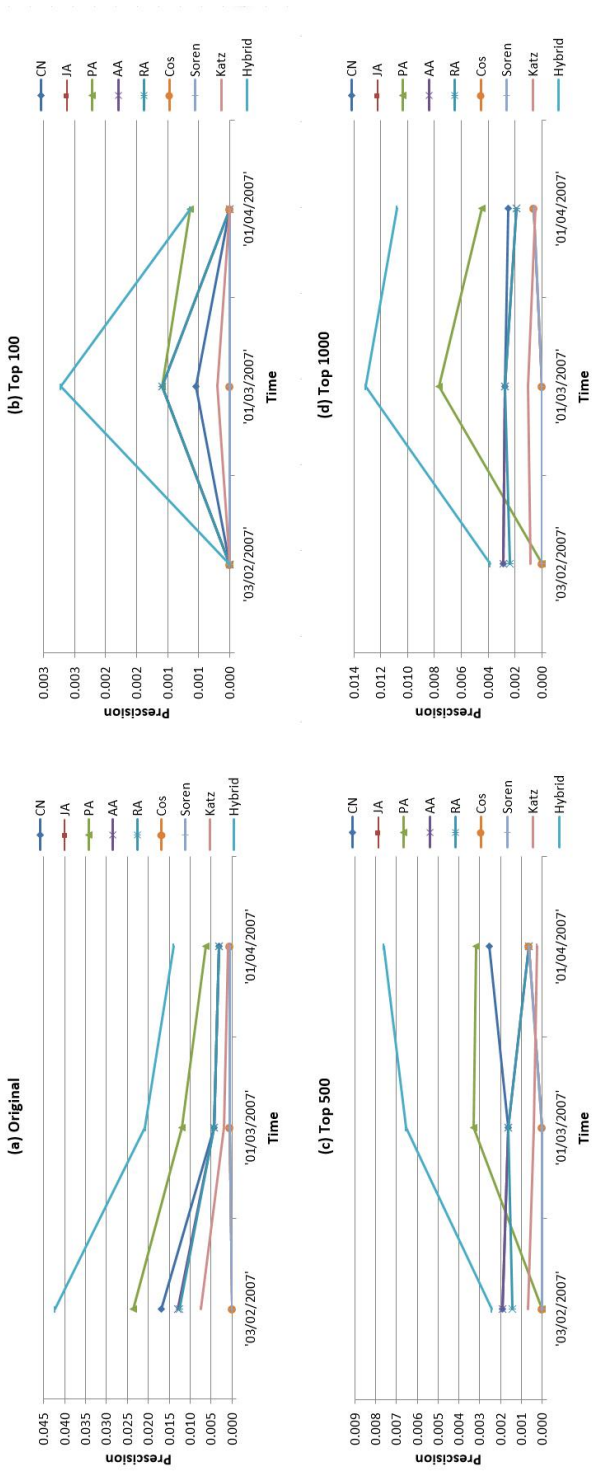


Figure 6.18: Flickr Monthly Sliding Window Prediction Recall Result

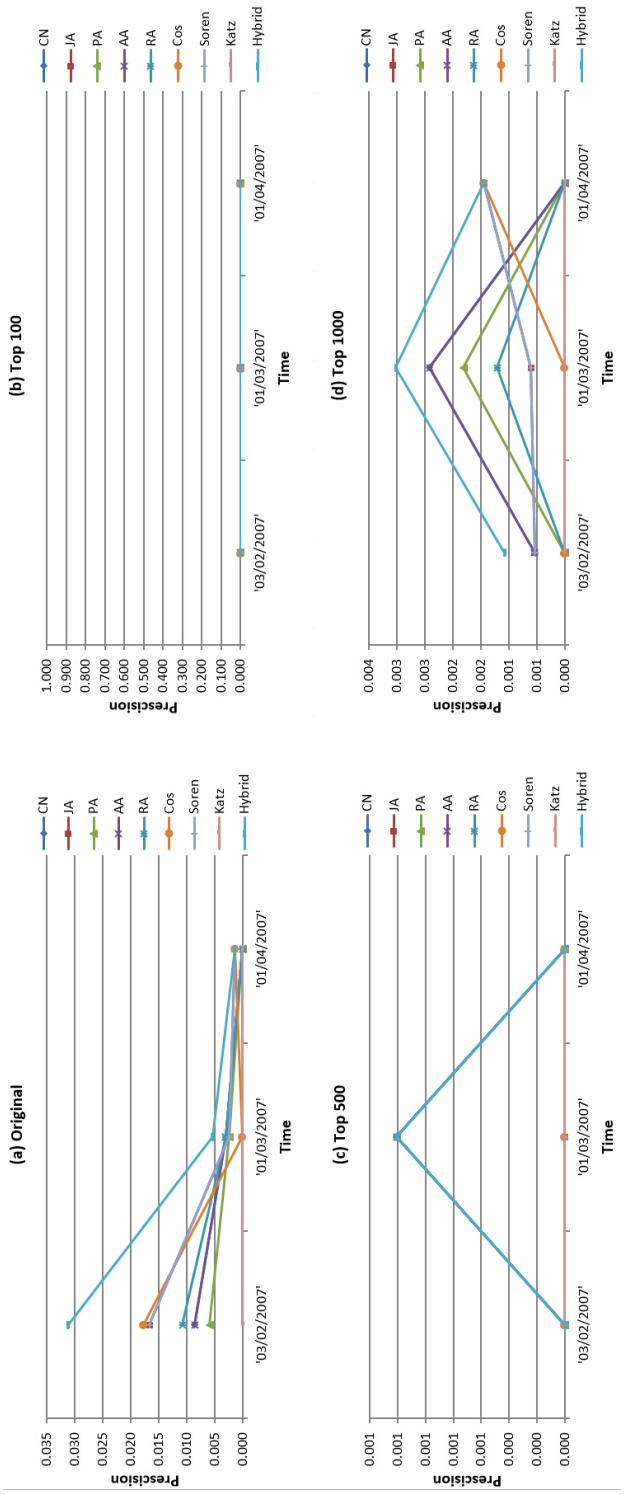


Figure 6.19: Flickr Monthly Growing Window Prediction Recall Result

Table 6.6: Flickr Prediction Average Precision

		Method	Original(std dev)	Top 50(std dev)	Top 100(std dev)	Top 500(std dev)	Top 1000(std dev)
Weekly	Slide	CN	0.0018 (0.0020)	0.0010 (0.0050)	0.0006 (0.0020)	0.0020(0.0020)	N/A
	Grow		0.0024 (0.0060)	0.0010 (0.0050)	0.0010 (0.0030)	**0.0030** (0.0070)	N/A
Monthly	Slide		0.0034 (0.0005)	N/A	0.0030 (0.0050)	0.0070 (0.0009)	0.0050 (0.0008)
	Grow		0.0014 (0.0010)	N/A	0.0000 (0.0000)	0.0010 (0.0020)	**0.0020** (0.0020)
Weekly	Slide	JA	0.0006 (0.0020)	0.0010 (0.0050)	0.0004 (0.0020)	0.0005 (0.0009)	N/A
	Grow		0.0006 (0.0020)	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)	N/A
Monthly	Slide		0.0003 (0.0002)	N/A	0.0000 (0.0000)	0.0007 (0.0009)	0.0003 (0.0005)
	Grow		**0.0022** (0.0004)	N/A	0.0000 (0.0000)	0.0000(0.0000)	0.0010 (0.0004)
Weekly	Slide	PA	**0.0094** (0.0177)	**0.0263** (0.0599)	**0.0175** (0.0356)	**0.0108** (0.0203)	N/A
	Grow		0.0024 (0.0054)	**0.0013** (0.0048)	0.0019 (0.0053)	0.0025 (0.0059)	N/A
Monthly	Slide		**0.0070** (0.0025)	N/A	0.0100 (0.0082)	**0.0073** (0.0052)	**0.0070** (0.0057)
	Grow		0.0011 (0.0009)	N/A	0.0000 (0.0000)	**0.0013**(0.0019)	0.0010 (0.0014)
Weekly	Slide	AA	0.0025 (0.0030)	0.0090 (0.0160)	0.0070 (0.0140)	0.0030 (0.0050)	N/A
	Grow		0.0023 (0.0050)	0.0000 (0.0000)	**0.0020** (0.0050)	**0.0030** (0.0070)	N/A
Monthly	Slide		0.0031 (0.0006)	N/A	**0.0070** (0.0090)	0.0050 (0.0020)	0.0050 (0.0010)
	Grow		0.0014 (0.0010)	N/A	0.0000 (0.0000)	0.0010 (0.0020)	0.0010 (0.0020)
Weekly	Slide	RA	0.0025 (0.0030)	0.0009 (0.0160)	0.0070 (0.0140)	0.0030 (0.0050)	N/A
	Grow		**0.0026** (0.0060)	0.0000 (0.0000)	0.0010 (0.0050)	**0.0030** (0.0070)	N/A
Monthly	Slide		0.0031 (0.0006)	N/A	**0.0070** (0.0090)	0.0050 (0.0020)	0.0040 (0.0009)
	Grow		0.0016 (0.0010)	N/A	0.0000 (0.0000)	0.0010 (0.0020)	0.0007 (0.0009)
Weekly	Slide	Cos	0.0006 (0.0008)	0.0010 (0.0050)	0.0006 (0.0020)	0.0005 (0.0009)	N/A
	Grow		0.0003 (0.0009)	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)	N/A
Monthly	Slide		0.0003 (0.0002)	N/A	0.0000 (0.0000)	0.0007 (0.0009)	0.0003 (0.0005)
	Grow		0.0015 (0.0010)	N/A	0.0000 (0.0000)	0.0000 (0.0000)	0.0007 (0.0009)
Weekly	Slide	Soren	0.0008 (0.0078)	0.0010 (0.0050)	0.0006 (0.0020)	0.0005 (0.0009)	N/A
	Grow		0.0006 (0.0020)	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)	N/A
Monthly	Slide		0.0003 (0.0002)	N/A	0.0000 (0.0000)	0.0007 (0.0009)	0.0003 (0.0005)
	Grow		**0.0022** (0.0004)	N/A	0.0000 (0.0000)	0.0000 (0.0000)	0.0010 (0.0005)
Weekly	Slide	Katz	0.0021 (0.0040)	0.0040 (0.0080)	0.0030 (0.0060)	0.0030 (0.0050)	N/A
	Grow		0.0024 (0.0060)	0.0010 (0.0050)	**0.0020** (0.0050)	**0.0030** (0.0070)	N/A
Monthly	Slide		0.0036 (0.0010)	N/A	0.0030 (0.0050)	0.0050 (0.0020)	0.0040 (0.0010)
	Grow		0.0015 (0.0010)	N/A	0.0000 (0.0000)	0.0010 (0.0020)	**0.0020** (0.0020)
Weekly	Slide	Hybrid	*0.0123 (0.0226)*	*0.0375 (0.0707)*	*0.0256 (0.0480)*	*0.0149 (0.0271)*	N/A
	Grow		*0.0039 (0.0098)*	*0.0013 (0.0048)*	*0.0038 (0.0105)*	*0.0036 (0.0089)*	N/A
Monthly	Slide		*0.0130 (0.0039)*	N/A	*0.0180 (0.0216)*	*0.0193 (0.0066)*	*0.0163 (0.0065)*
	Grow		*0.0040 (0.0020)*	N/A	*0.0000 (0.0000)*	*0.0013 (0.0019)*	*0.0030 (0.0014)*
Weekly	Slide	Increase	31%	43%	46%	38%	N/A
	Grow		50%	0%	90%	20%	N/A
Monthly	Slide		86%	N/A	157%	164%	133%
	Grow		82%	N/A	0%	0%	50%

Note: This table summarized the average precision result for Flickr network. Each row represented the average precision for each prediction methods with weekly and monthly prediction in both sliding and grow scenarios with different number we predicted. The numbers in bold are the best performed method amongst select methods in different prediction scenarios.

Table 6.7: Flickr Prediction Average Recall

		Method	Original(std dev)	Top 50(std dev)	Top 100(std dev)	Top 500(std dev)	Top 1000(std dev)
Weekly	Slide	CN	0.0024 (0.0044)	0.0001 (0.0003)	0.0001 (0.0003)	0.0012(0.0014)	N/A
	Grow		0.0015 (0.0023)	0.0000 (0.0000)	0.0000 (0.0000)	0.0005 (0.0012)	N/A
Monthly	Slide		0.0081 (0.0062)	N/A	0.0002 (0.0003)	0.0020 (0.0004)	0.0027 (0.0001)
	Grow		0.0039 (0.0036)	N/A	0.0000 (0.0000)	**0.0004** (0.0006)	**0.0010** (0.0010)
Weekly	Slide	JA	0.0006 (0.0009)	0.0001 (0.0003)	0.0001 (0.0003)	0.0005 (0.0008)	N/A
	Grow		0.0013 (0.0042)	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)	N/A
Monthly	Slide		0.0004 (0.0003)	N/A	0.0000 (0.0000)	0.0002 (0.0003)	0.0002 (0.0003)
	Grow		**0.0068** (0.0070)	N/A	0.0000 (0.0000)	0.0000(0.0000)	0.0009 (0.0004)
Weekly	Slide	PA	**0.0070** (0.0090)	**0.0007** (0.0011)	**0.0012** (0.0020)	**0.0037** (0.0047)	N/A
	Grow		**0.0017** (0.0026)	0.0000 (0.0000)	0.0000 (0.0000)	0.0006 (0.0011)	N/A
Monthly	Slide		**0.0140** (0.0072)	N/A	**0.0006** (0.0005)	**0.0022** (0.0015)	**0.0040** (0.0031)
	Grow		0.0028 (0.0023)	N/A	0.0000 (0.0000)	0.0004(0.0005)	0.0006 (0.0008)
Weekly	Slide	AA	0.0029 (0.0046)	0.0004 (0.0009)	0.0005 (0.0010)	0.0014 (0.0014)	N/A
	Grow		0.0015 (0.0023)	0.0000 (0.0000)	**0.0001** (0.0002)	**0.0007** (0.0012)	N/A
Monthly	Slide		0.0068 (0.0044)	N/A	0.0004 (0.0005)	0.0014 (0.0006)	0.0025 (0.0004)
	Grow		0.0039 (0.0036)	N/A	0.0000 (0.0000)	**0.0004** (0.0006)	0.0010 (0.0010)
Weekly	Slide	RA	0.0030 (0.0049)	0.0004 (0.0009)	0.0005 (0.0010)	0.0013 (0.0014)	N/A
	Grow		**0.0017** (0.0027)	0.0000 (0.0000)	0.0000 (0.0000)	0.0006 (0.0009)	N/A
Monthly	Slide		0.0067 (0.0041)	N/A	0.0004 (0.0005)	0.0012 (0.0004)	0.0023 (0.0003)
	Grow		0.0046 (0.0045)	N/A	0.0000 (0.0000)	**0.0004** (0.0006)	0.0004 (0.0006)
Weekly	Slide	Cos	0.0006 (0.0009)	0.0001 (0.0003)	0.0001 (0.0003)	0.0005 (0.0008)	N/A
	Grow		0.0010 (0.0034)	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)	N/A
Monthly	Slide		0.0004 (0.0003)	N/A	0.0000 (0.0000)	0.0002 (0.0003)	0.0002 (0.0003)
	Grow		0.0064 (0.0080)	N/A	0.0000 (0.0000)	0.0000 (0.0000)	0.0005(0.0007)
Weekly	Slide	Soren	0.0006 (0.0009)	0.0001 (0.0003)	0.0001 (0.0003)	0.0005 (0.0008)	N/A
	Grow		0.0013 (0.0042)	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)	N/A
Monthly	Slide		0.0004 (0.0003)	N/A	0.0000 (0.0000)	0.0007 (0.0009)	0.0002 (0.0003)
	Grow		**0.0068** (0.0070)	N/A	0.0000 (0.0000)	0.0000 (0.0000)	0.0005 (0.0007)
Weekly	Slide	Katz	0.0011 (0.0019)	0.0001 (0.0002)	0.0001 (0.0002)	0.0005 (0.0006)	N/A
	Grow		0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)	N/A
Monthly	Slide		0.0034 (0.0028)	N/A	0.0001 (0.0001)	0.0004 (0.0002)	0.0008 (0.0002)
	Grow		0.0000 (0.0000)	N/A	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)
Weekly	Slide	Hybrid	*0.0105 (0.0140)*	*0.0012 (0.0016)*	*0.0018 (0.0480)*	*0.0053 (0.0055)*	N/A
	Grow		*0.0032 (0.0063)*	*0.0000 (0.0000)*	*0.0001 (0.0004)*	*0.0008 (0.0014)*	N/A
Monthly	Slide		*0.0257 (0.0121)*	N/A	*0.0011 (0.0012)*	*0.0055 (0.0023)*	*0.0093 (0.0039)*
	Grow		*0.0127 (0.0131)*	N/A	*0.0000 (0.0000)*	*0.0004 (0.0006)*	*0.0018 (0.0008)*
Weekly	Slide	Increase	50%	71%	50%	43%	N/A
	Grow		88%	0%	0%	14%	N/A
Monthly	Slide		84%	N/A	83%	150%	133%
	Grow		87%	N/A	0%	0%	80%

Note: This table summarized the average recall result for Flickr network. Each row represented the average recall for each prediction methods with weekly and monthly prediction in both sliding and grow scenarios with different number we predicted. The numbers in bold are the best performed method amongst select methods in different prediction scenarios.

6.4.5 Twitter Network

Fig 6.20 to Fig 6.23 depict the precision and recall results for the Twitter network. As shown in Fig 6.20, the best prediction results for growing window is from hybrid model for the Top 10 link prediction scenario with a precision of 0.4. For sliding window prediction, the best result also obtained in Top 10 links prediction with precision of 0.2 (Fig 6.21). Different from Flickr network, the preferential attachment prediction precision results on Twitter network are not perform well. Katz prediction method is the best performed prediction method on average as it gives best precision when predicting top 50 and 100 links for both sliding and growing settings.

In both sliding and growing window scenarios, the decreasing trend of prediction precision is observed. This is due to the decrease number of new links between each steps as shown in Table 6.4.5. We have a better chance to do a correct prediction when many links appear. However, in the Sliding Window prediction on 18/10/2009, there are not many new links formed but the precision is high especially when we predict Top 10 links. So the number of new links could influence the prediction accuracy but the effect is not the determining factor.

Table 6.9 and Table 6.10 indicate the average prediction precision and recall results. On average, our hybrid model improve the precision by at least 33% and most 99%. The recall improvement of our hybrid model is in the range of 22% to 187%.

Time (2009)	15/10	16/10	17/10	18/10	19/10	20/10
New Links	301	634	193	121	198	28

Table 6.8: Twitter New Links Number For Each Step

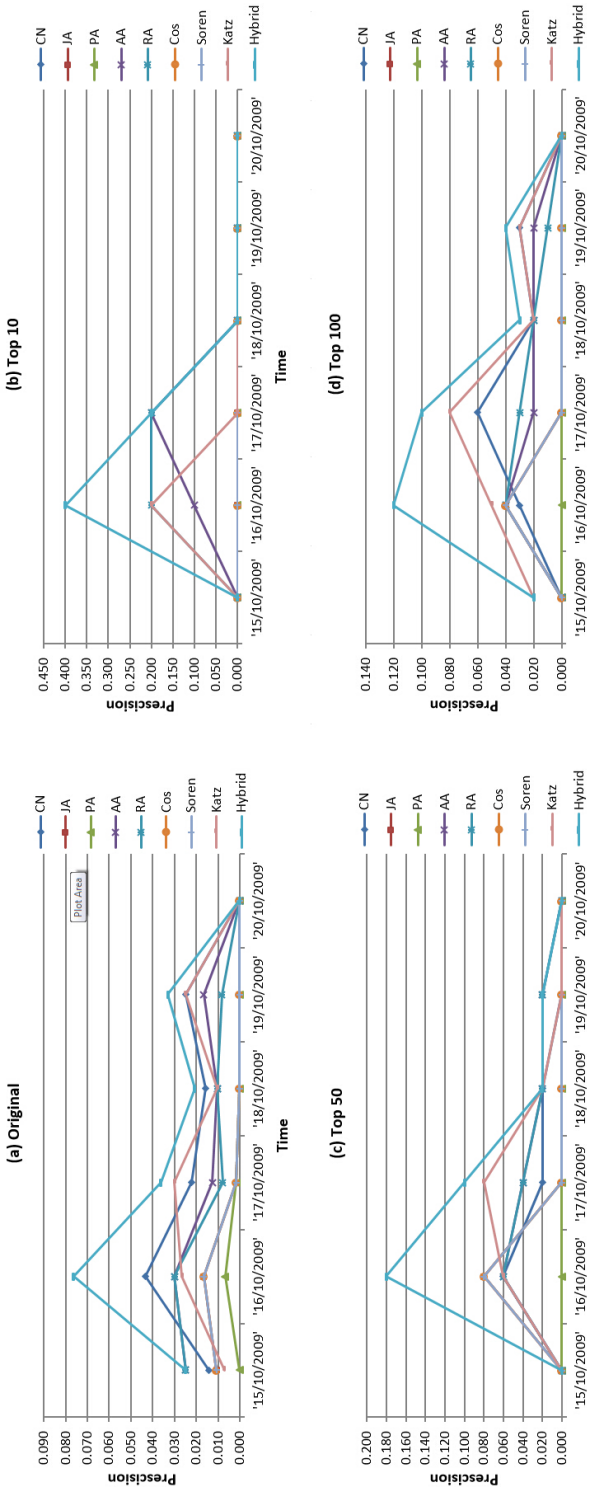


Figure 6.20: Twitter Daily Growing Window Prediction Precision Result

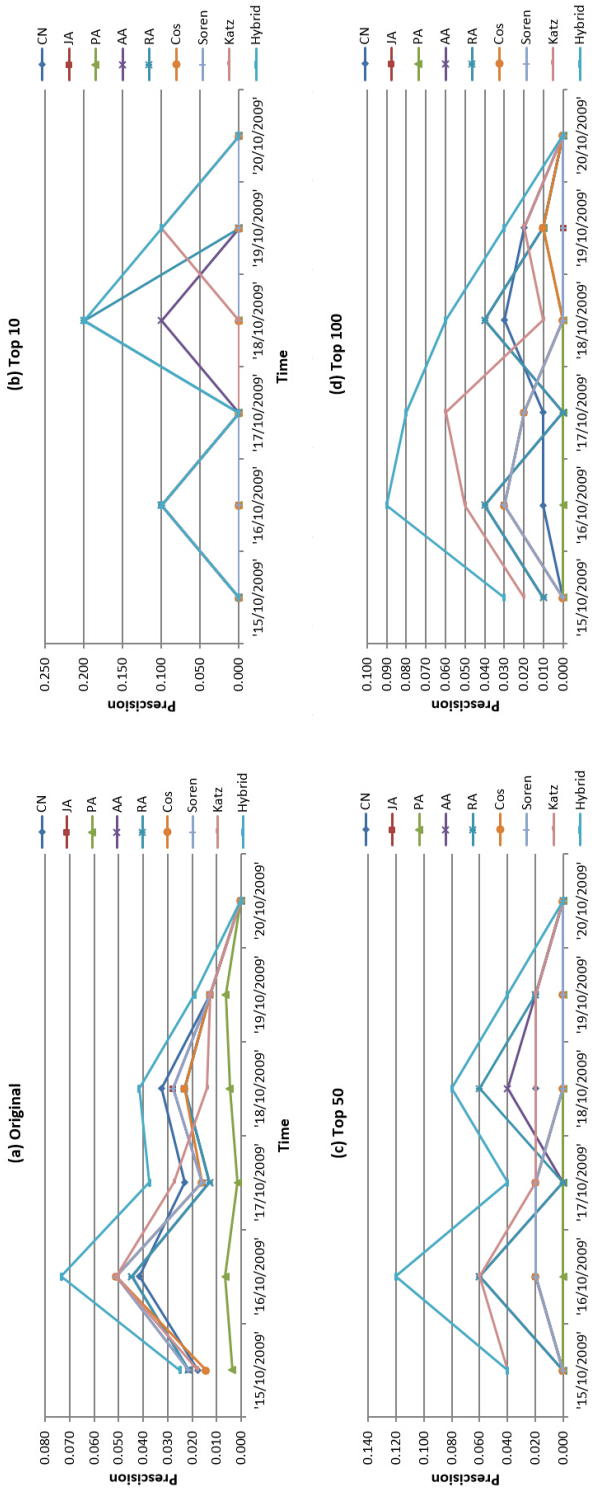


Figure 6.21: Twitter Daily Sliding Window Prediction Precision Result

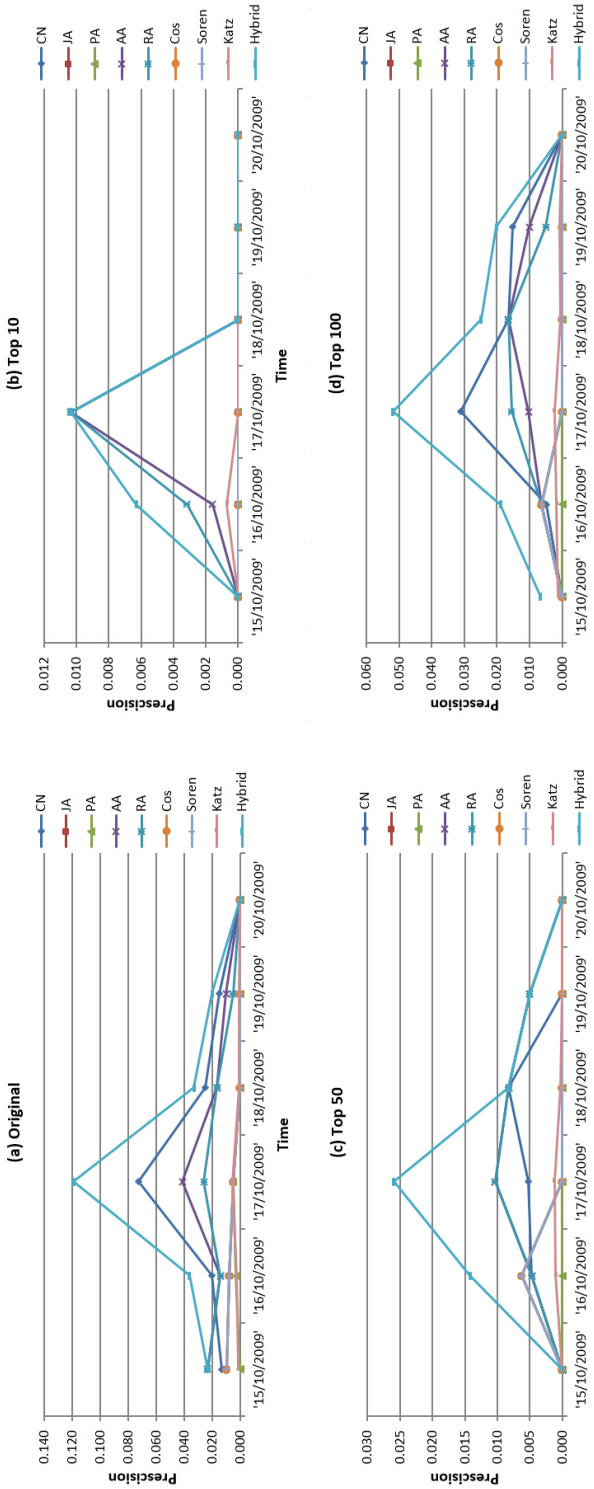


Figure 6.22: Twitter Daily Growing Window Prediction Recall Result

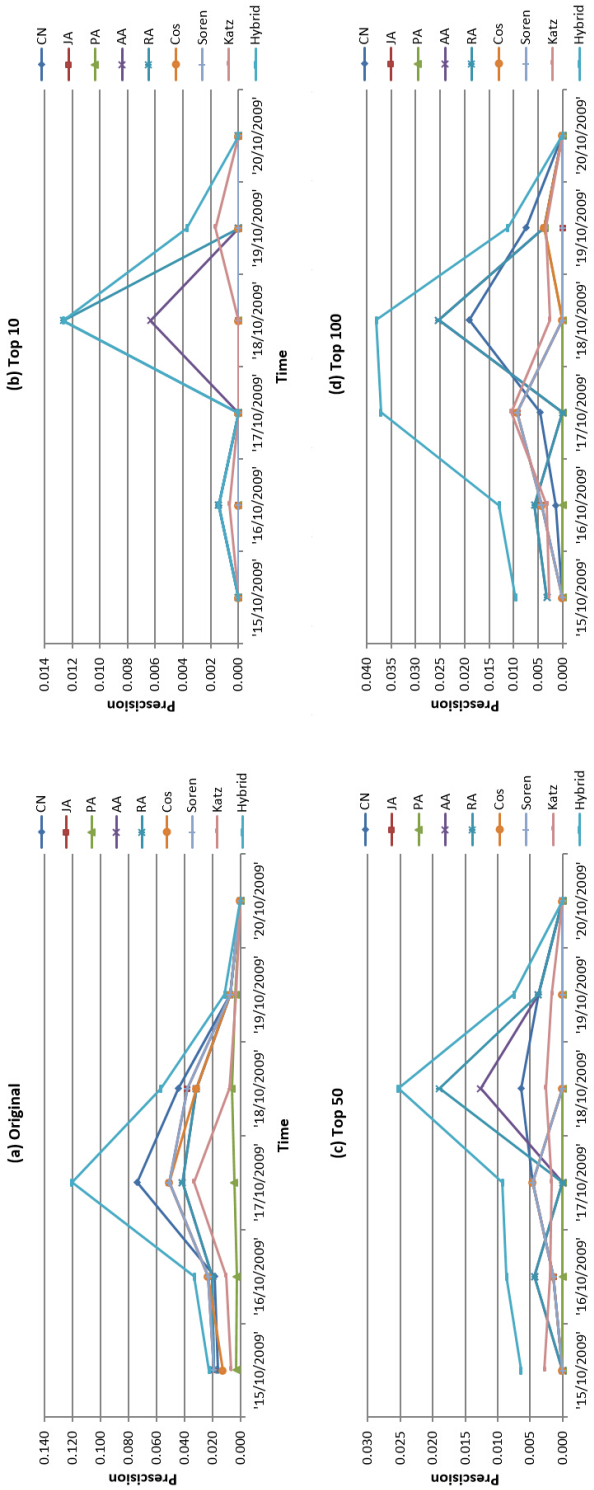


Figure 6.23: Twitter Daily Sliding Window Prediction Recall Result

Table 6.9: Twitter Prediction Average Precision

		Method	Original(std dev)	Top 10(std dev)	Top 50(std dev)	Top 100(std dev)
Daily	Slide	CN	0.0212 (0.0134)	0.0167 (0.0373)	0.0133 (0.0094)	0.0117(0.0107)
	Grow		**0.0200** (0.0130)	0.0000 (0.0000)	0.0167 (0.0213)	0.0233 (0.0205)
Daily	Slide	JA	**0.0215** (0.0158)	0.0000 (0.0000)	0.0067 (0.0094)	0.0083 (0.0121)
	Grow		0.0048 (0.0065)	0.0000 (0.0000)	0.0133 (0.0298)	0.0067 (0.0149)
Daily	Slide	PA	0.0037 (0.0024)	0.0000 (0.0000)	0.0000 (0.0000)	0.0017 (0.0037)
	Grow		0.0014 (0.0024)	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)
Daily	Slide	AA	0.0191 (0.0137)	0.0333 (0.0471)	0.0200 (0.0231)	0.0167 (0.0170)
	Grow		0.0157 (0.0098)	0.0500 (0.0768)	0.0233 (0.0213)	0.0163 (0.0137)
Daily	Slide	RA	0.0191 (0.0137)	**0.0500** (0.0764)	0.0233 (0.0269)	0.0167 (0.0170)
	Grow		0.0136 (0.0104)	**0.0667** (0.0943)	0.0233 (0.0213)	0.0167 (0.0149)
Daily	Slide	Cos	0.0195 (0.0157)	0.0000 (0.0000)	0.0067 (0.0094)	0.0083 (0.0121)
	Grow		0.0048 (0.0065)	0.0000 (0.0000)	0.0133 (0.0298)	0.0067 (0.0149)
Daily	Slide	Soren	**0.0215** (0.0158)	0.0000 (0.0000)	0.0067 (0.0094)	0.0083 (0.0121)
	Grow		0.0048 (0.0065)	0.0000 (0.0000)	0.0133 (0.0298)	0.0067 (0.0149)
Daily	Slide	Katz	0.0205 (0.0159)	0.0333 (0.0471)	**0.0267** (0.0189)	**0.0267** (0.0213)
	Grow		0.0165 (0.0112)	0.0333 (0.0745)	**0.0267** (0.0320)	**0.0333** (0.0256)
Daily	Slide	Hybrid	*0.0327 (0.0227)*	*0.0667 (0.0745)*	*0.0533 (0.0377)*	*0.0483 (0.0313)*
	Grow		*0.0319 (0.0231)*	*0.1000 (0.1528)*	*0.0533 (0.0660)*	*0.0516 (0.0434)*
Daily	Slide	Increase	52%	33%	99%	81%
	Grow		60%	50%	99%	55%

Table 6.10: Twitter Prediction Average Recall

		Method	Original(std dev)	Top 10(std dev)	Top 50(std dev)	Top 100(std dev)
Daily	Slide	CN	**0.0267** (0.0250)	0.0002 (0.0005)	0.0027 (0.0024)	0.0054(0.0066)
	Grow		**0.0244** (0.0229)	0.0000 (0.0000)	0.0030 (0.0032)	0.0113 (0.0110)
Daily	Slide	JA	0.0231 (0.0173)	0.0000 (0.0000)	0.0010 (0.0017)	0.0023 (0.0035)
	Grow		0.0038 (0.0041)	0.0000 (0.0000)	0.0011 (0.0024)	0.0011 (0.0024)
Daily	Slide	PA	0.0035 (0.0019)	0.0000 (0.0000)	0.0000 (0.0000)	0.0006(0.0014)
	Grow		0.0014 (0.0021)	0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)
Daily	Slide	AA	0.0200 (0.0139)	0.0013 (0.0023)	0.0034 (0.0045)	**0.0063** (0.0087)
	Grow		0.0176 (0.0128)	0.0020 (0.0038)	**0.0047** (0.0039)	**0.0072** (0.0059)
Daily	Slide	RA	0.0200 (0.0139)	**0.0023** (0.0046)	**0.0045** (0.0067)	**0.0063** (0.0087)
	Grow		0.0142 (0.0092)	**0.0023** (0.0038)	**0.0047** (0.0039)	**0.0072** (0.0067)
Daily	Slide	Cos	0.0210 (0.0169)	0.0000 (0.0000)	0.0010 (0.0017)	0.0029 (0.0034)
	Grow		0.0038 (0.0041)	0.0000 (0.0000)	0.0011 (0.0024)	0.0011 (0.0024)
Daily	Slide	Soren	0.0231 (0.0173)	0.0000 (0.0000)	0.0010 (0.0017)	0.0023 (0.0035)
	Grow		0.0038 (0.0041)	0.0000 (0.0000)	0.0011 (0.0024)	0.0011 (0.0024)
Daily	Slide	Katz	0.0103 (0.0108)	0.0004 (0.0006)	0.0018 (0.0009)	0.0037 (0.0032)
	Grow		0.0018 (0.0019)	0.0001 (0.0003)	0.0004 (0.0005)	0.0011 (0.0008)
Daily	Slide	Hybrid	*0.0407 (0.0399)*	*0.0030 (0.0045)*	*0.0095 (0.0077)*	*0.0181 (0.0143)*
	Grow		*0.0387 (0.0378)*	*0.0028 (0.0041)*	*0.0089 (0.0091)*	*0.0204 (0.0164)*
Daily	Slide	Increase	52%	30%	111%	187%
	Grow		59%	22%	89%	183%

6.4.6 Methods Weight

Fig 6.24 to Fig 6.37 depict the weight for each method we obtained in different experimental settings. As we can see from the figures, the weight for each method changes everytime when we change the time window. The weight shows how much one method (or rule to be consistent with section 6.2) contributes to the combination of predictors. Thus, these figures show the changing of the rules that the network evolution follows. The weight rank of a method is not necessarily in propotion to the prediction accuracy rank of the method among all the selected methods. In another word, a method that outperforms all the other methods does not implicit the method will be allocated the largest weight in the combination. Otherwise, the hybrid model would be meaningless as one can simply assign the weights for each method in proportion to the prediction accuracy result given by the method alone. For example, in Fig 6.2 (d), the best performed weight among the eight combined methods is Katz at first 5 window steps and AA at the following 4 steps. However, in Fig 6.24 (d), we did not observe that the two methods had the largest weights in the combination and the hybrid result is still better than all the eight selected methods.

The experimental variable N, which is a number of links that is predicted, is another factor which causes the difference in weights. Within each figure, the weights are changing differently for the four sub-figures due to the fact that we are predicting different number of new links. But we can still see the weight changing trends of the four sub-figures are similar to each other. Also, different window sizes and step sizes also lead to the variation of weights. Thus, to solve the real world problems with the hybrid model, we need to choose the appropriate experiment settings which fits the problem well.

From the perspective of changes in methods weights, which reflect the net-

work evolution, we can see the difference between Facebook and PWr networks. The Facebook one, although weights for methods vary depending on experimental setting, is evolving mainly following rules described by five methods: Katz, RA, PA, CN and AA. As we observed the majority of the weight are allocated to these methods. The weights for different window steps are not changing dramatically which means the Facebook network are evolution is rather stable. Different from Facebook, the changes in weights for PWr network are relative irregular (Fig 6.29 and Fig 6.31). We did not find similar patterns between the four sub-figures in each figure. The weights are quite sensitive to the number of links we predicted. It is because the PWr network evolution is very dynamic and changes rapidly. The evolving rule changes so fast, that for each window step we get very different weights. Besides, in Fig 6.28 and Fig 6.30, there are many window steps in which weight for RA method dominates over all other methods. Refering to Fig 6.8 and Fig 6.10, the weight of RA becomes dominant when all the prediction methods, including RA, perform very poorly even with the precision result down to 0 in some window steps. It can be explained by looking at the Equation 6.1. The optimization process is trying to find the best weight vector such that the distance between weighted combined prediction similarity score matrix and the new link matrix, which measured by Euclidean Norm, is as small as possible. When a method gives prediction precision 0, none of the predicted links is correct. In this case, the prediction result matrix that contains smaller similarity score are more close to the new link matrix and thus will be given a higher weight. RA is the method that gives the smallest similarity score and thus been given a larger weight when all the methods perform poorly. In Fig. 6.32 and Fig. 6.33, the PA was given the highest weight which indicate how the Flickr network evolves. The PA method also performs better than the methods other than our hybrid method. But when we look at the Fig. 6.32 and Fig. 6.34, the RA

and PA together were the rules which the network evolution follows. So the way how network changes also relates to the methods we adopt for network analysis. For Twitter network, the weight for each method is not steady as window moves. No method could be dominant in all window steps as depicted in Fig.6.36 and Fig.6.37. It is also a network with rapidly changing dynamics and in this way similar to the PWr network.

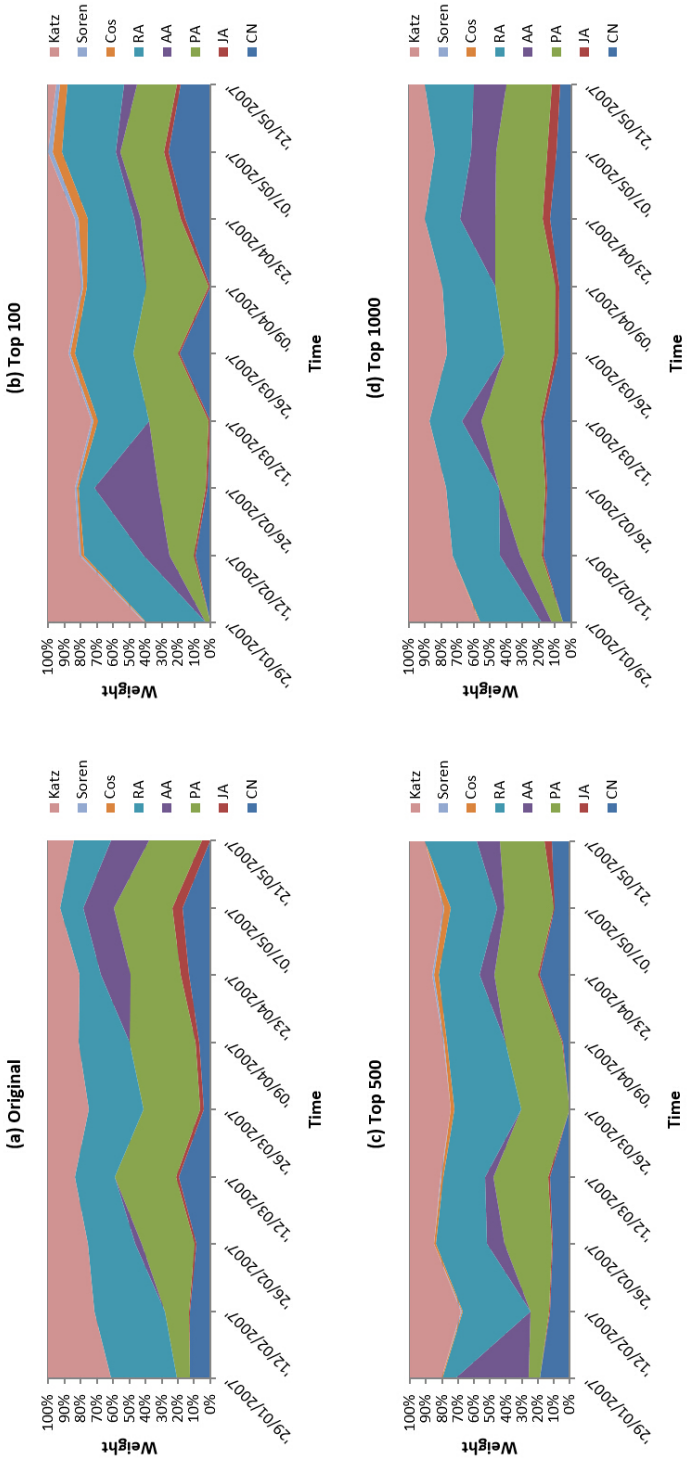


Figure 6.24: Facebook Monthly Growing Window Method Weight

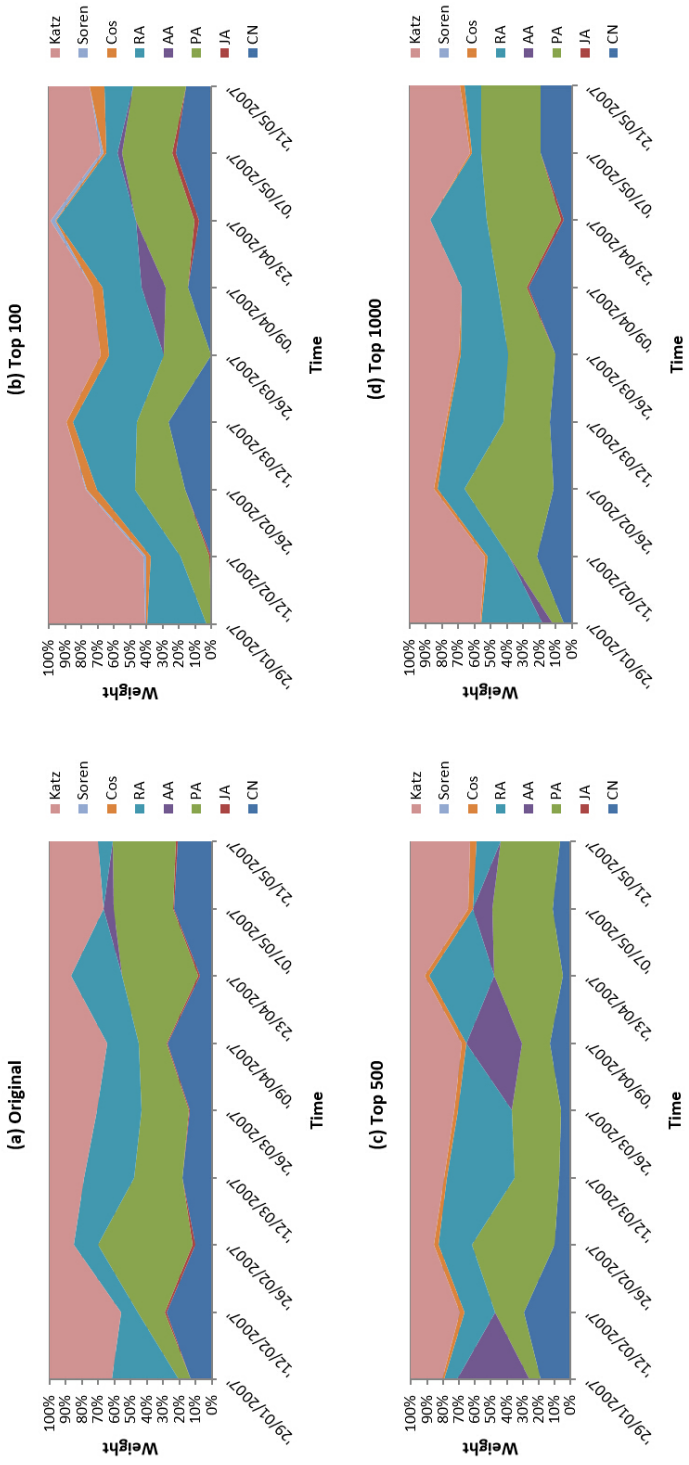


Figure 6.25: Facebook Monthly Sliding Window Method Weight

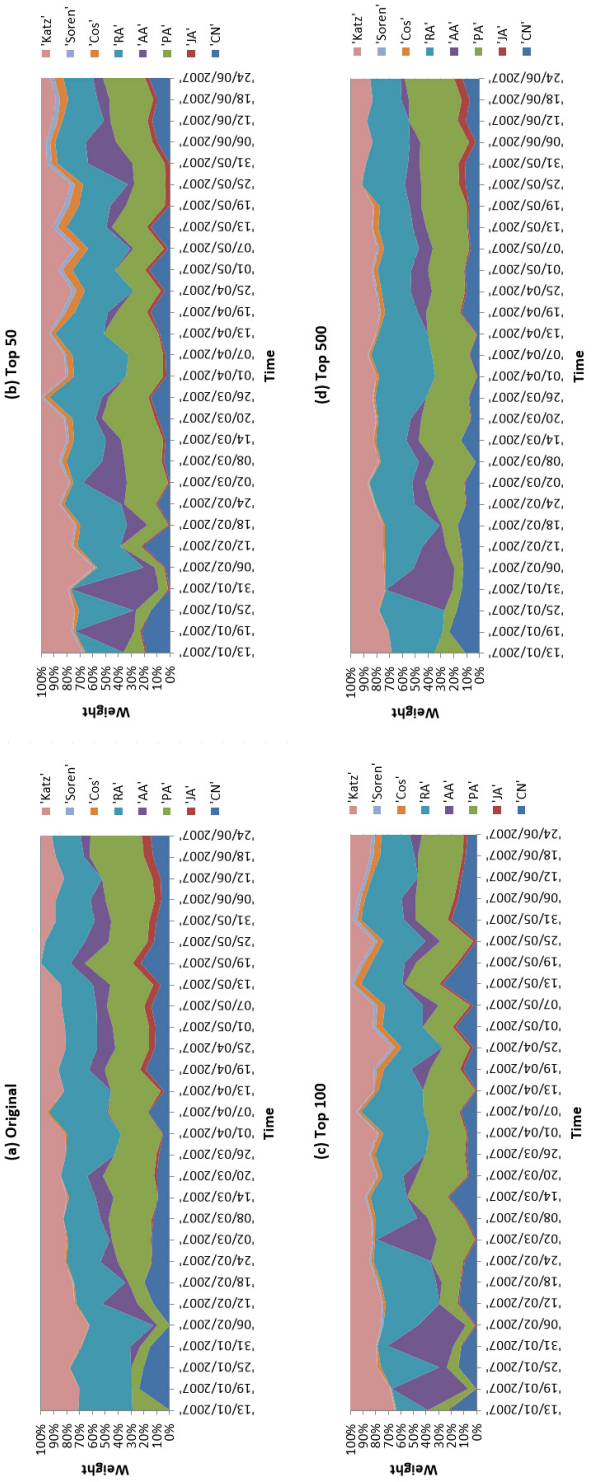


Figure 6.26: Facebook Weekly Growing Window Method Weight

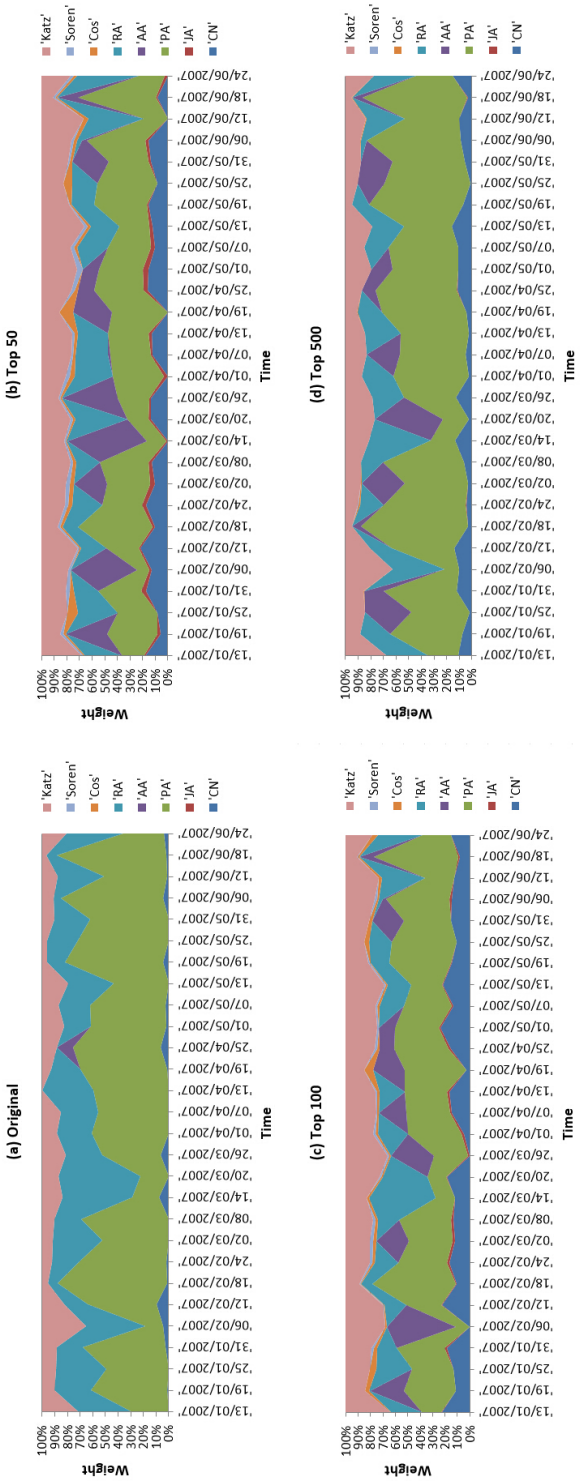


Figure 6.27: Facebook Weekly Sliding Window Method Weight

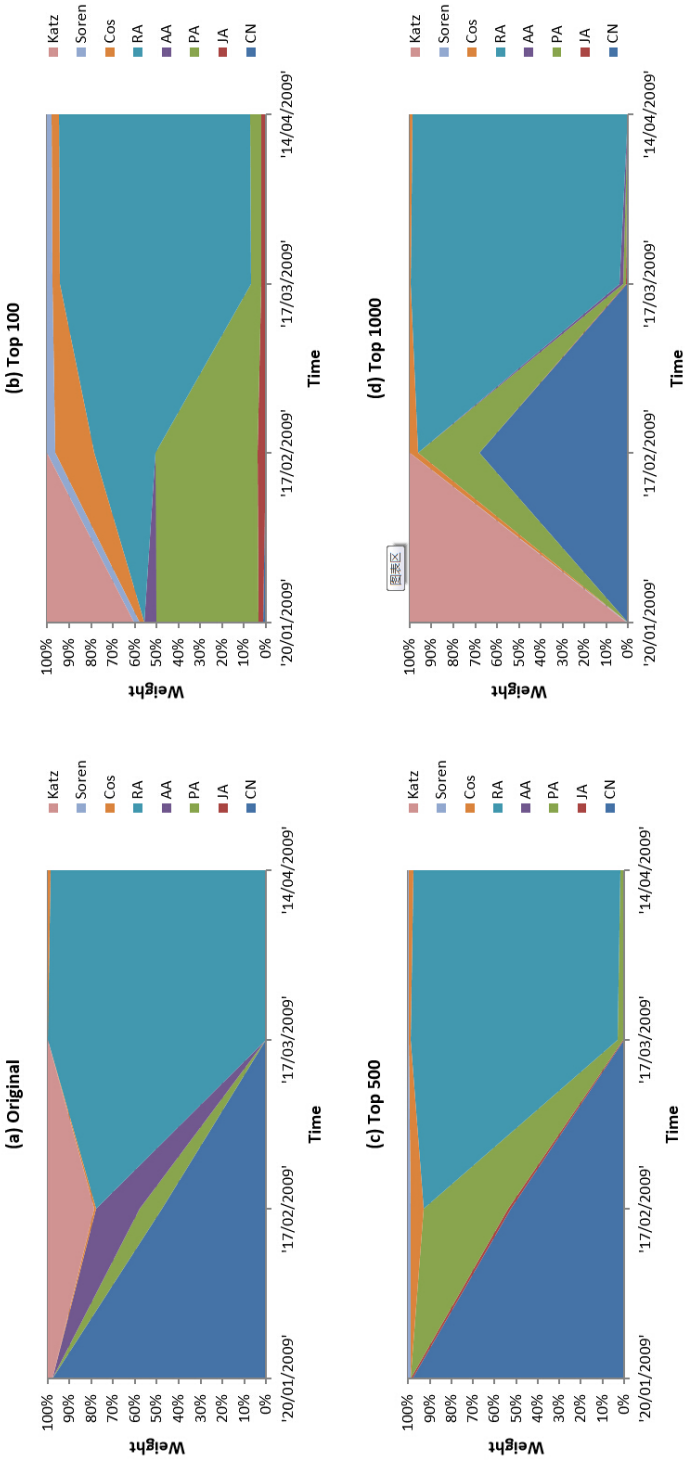


Figure 6.28: PWr Monthly Growing Window Method Weight

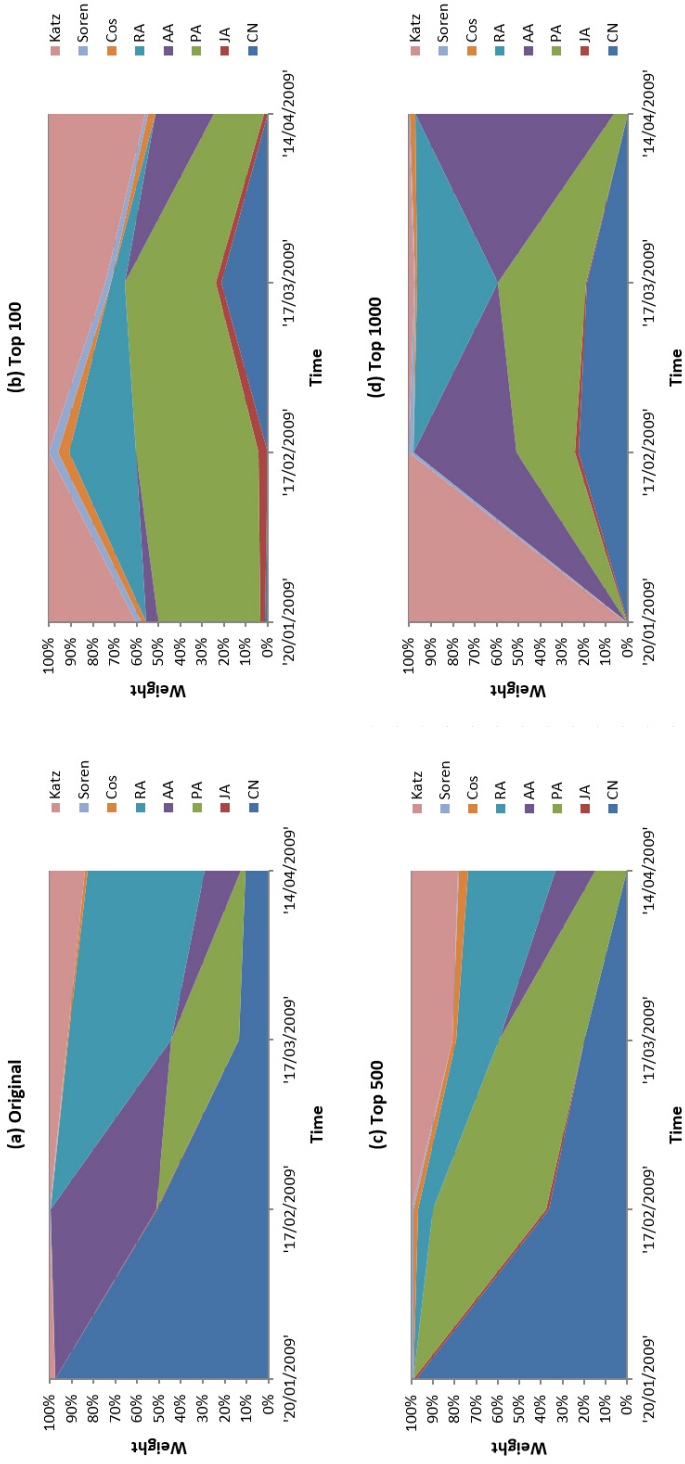


Figure 6.29: PWr Monthly Sliding Window Method Weight

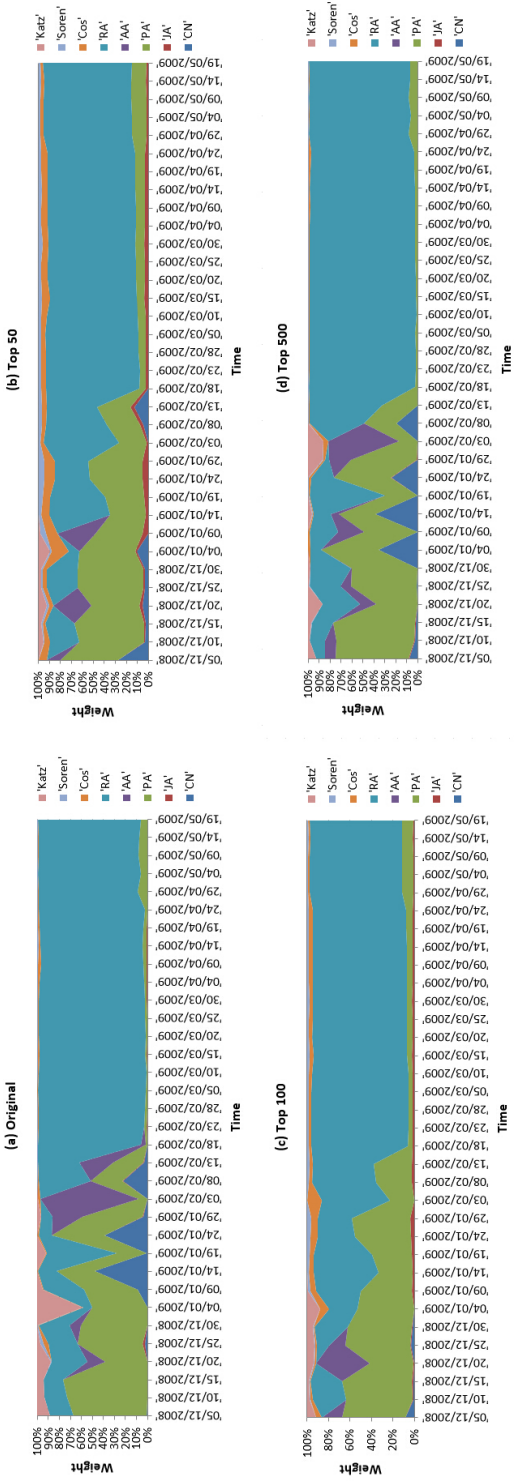


Figure 6.30: PWr Weekly Growing Window Method Weight

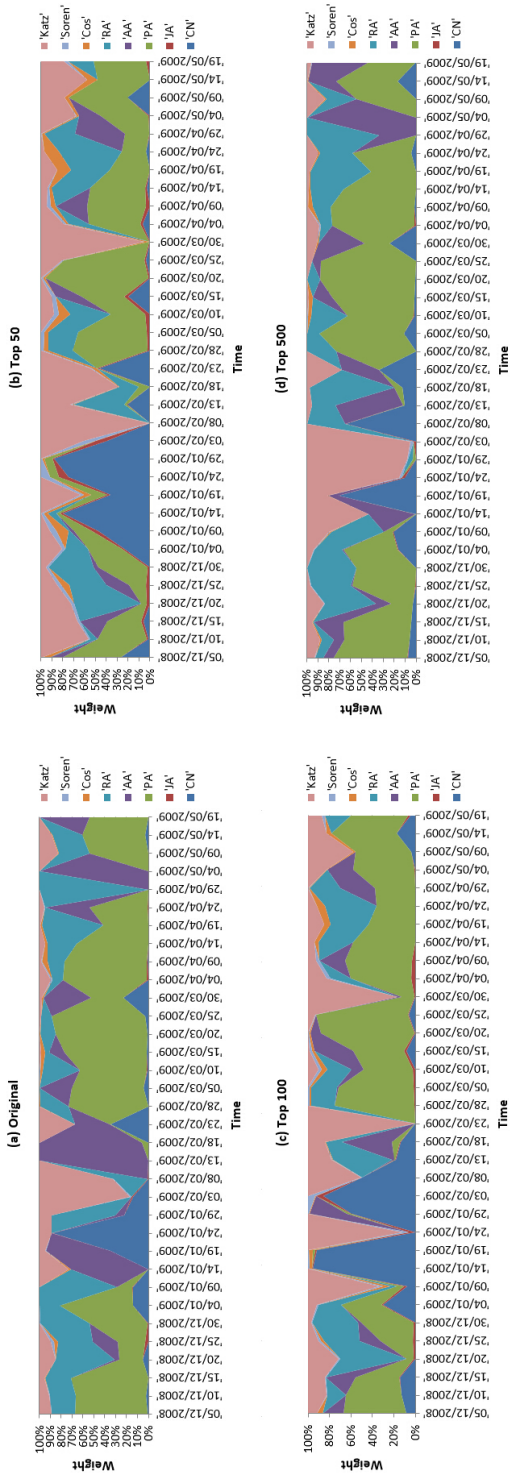


Figure 6.31: PWr Weekly Sliding Window Method Weight

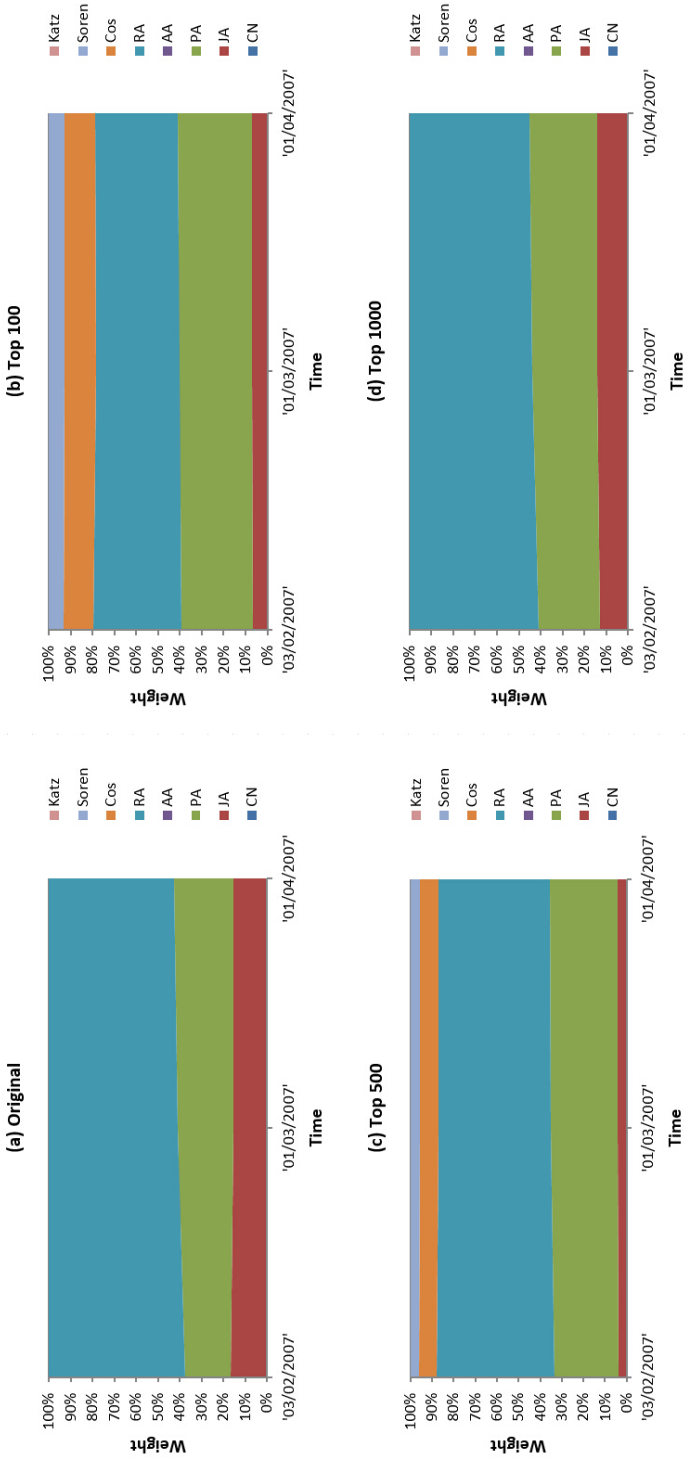


Figure 6.32: Flickr Monthly Growing Window Method Weight

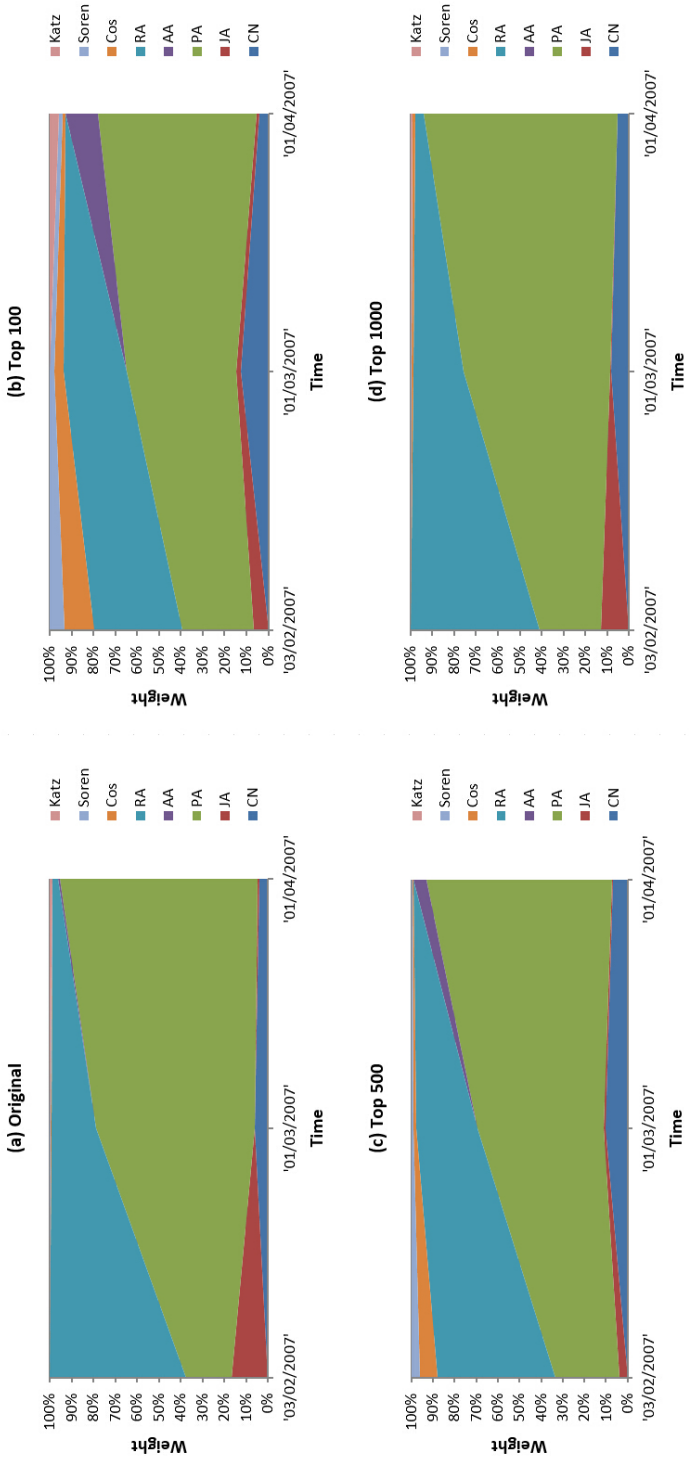


Figure 6.33: Flickr Monthly Sliding Window Method Weight

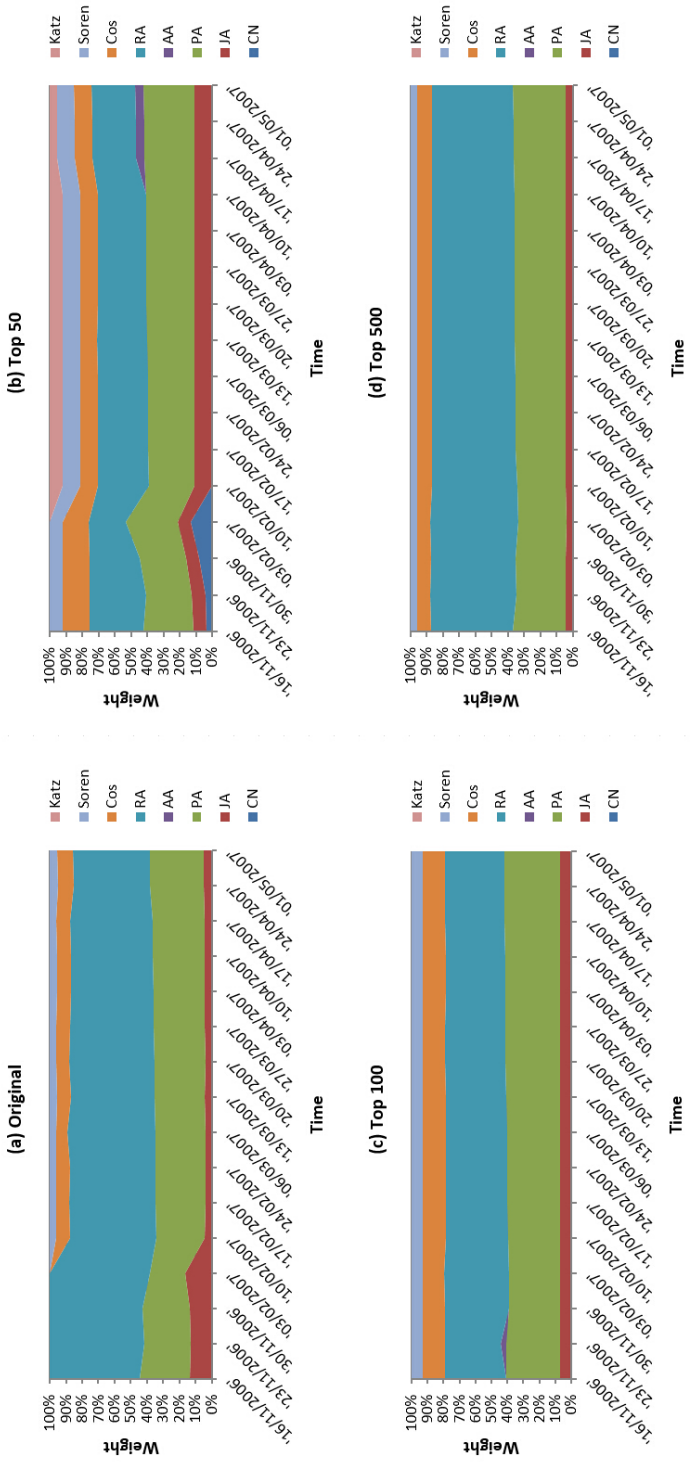


Figure 6.34: Flickr Weekly Growing Window Method Weight

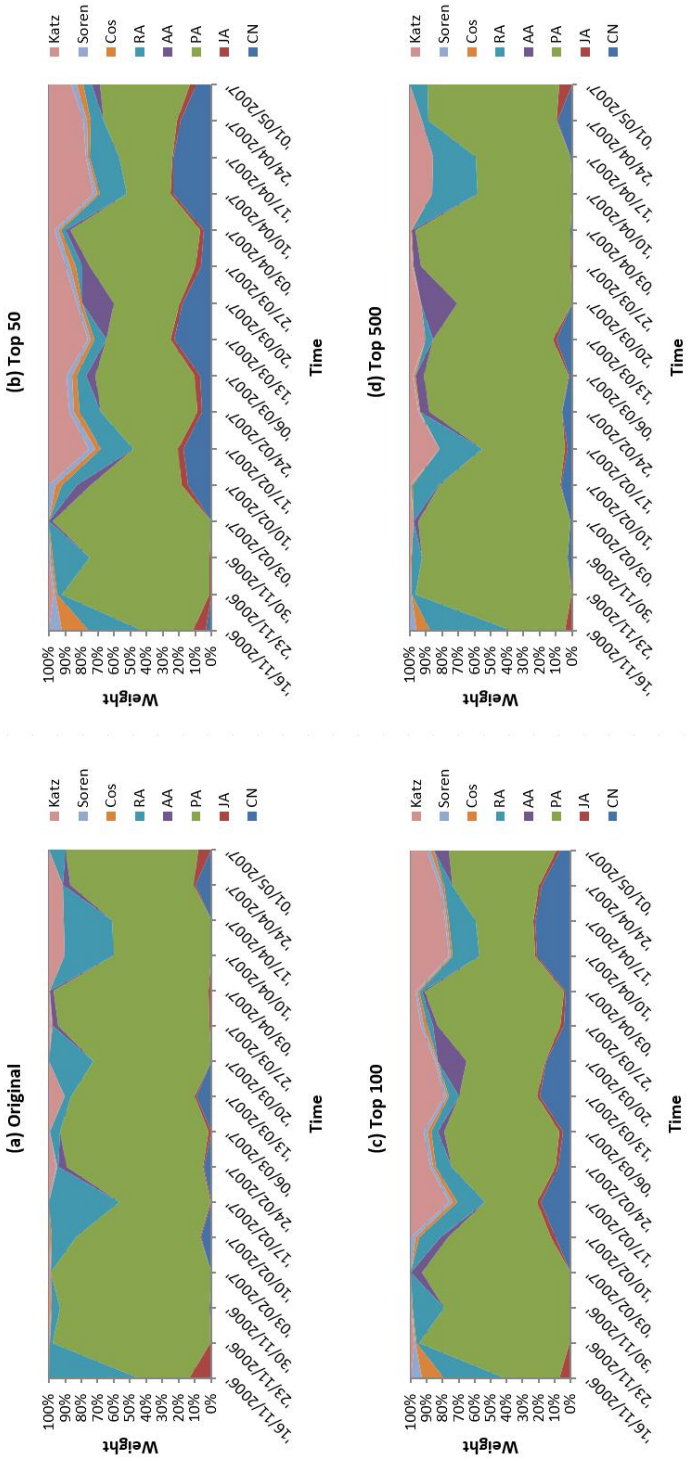


Figure 6.35: Flickr Weekly Sliding Window Method Weight

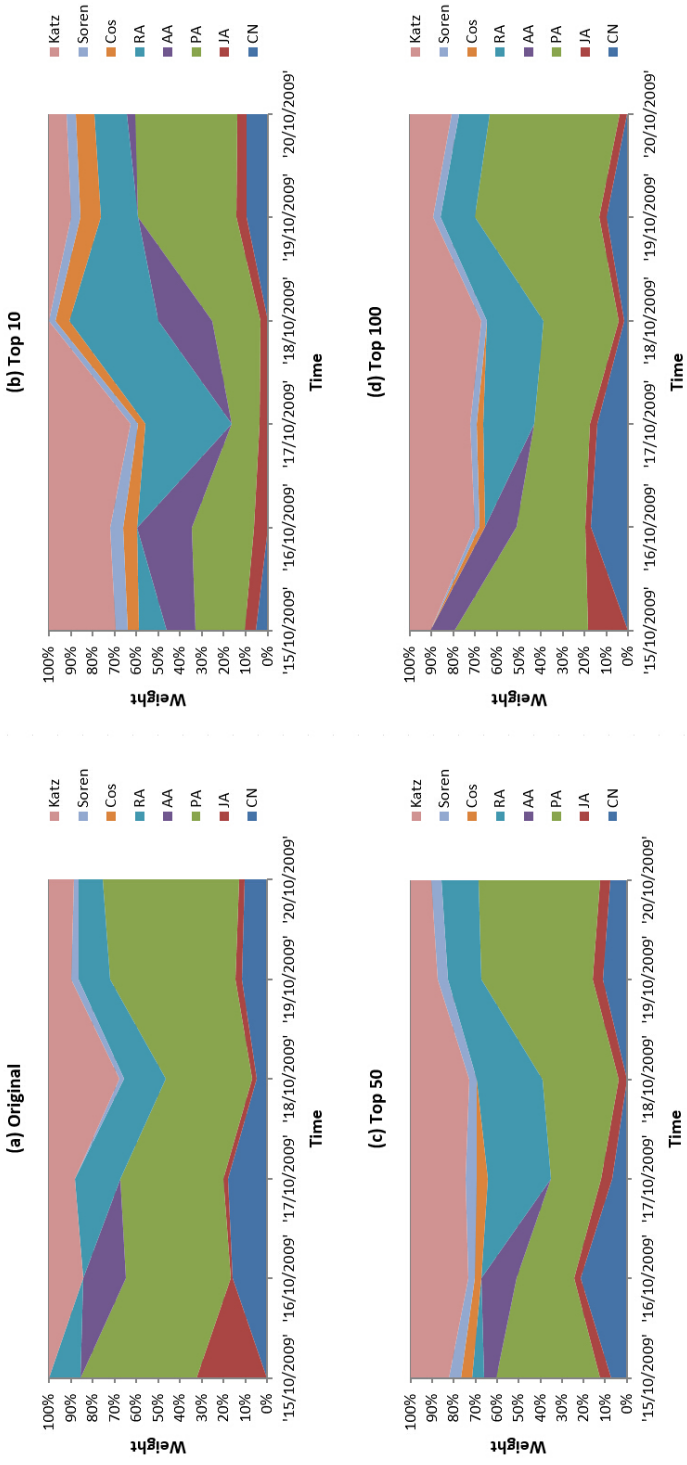


Figure 6.36: Twitter Daily Growing Window Method Weight

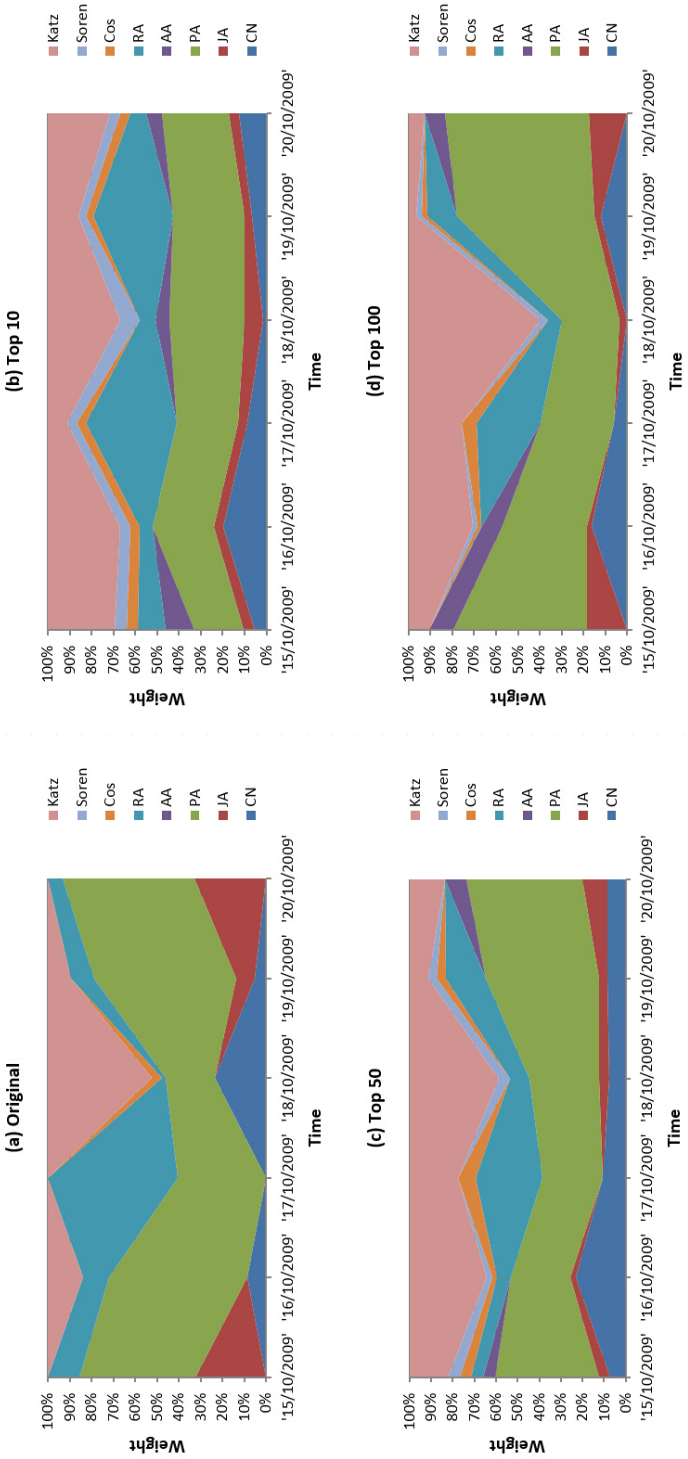


Figure 6.37: Twitter Daily Sliding Window Method Weight

6.4.7 Dataset Network Topology

The performance of hybrid model varies on the analysed networks. To understand the reason behind it, we explored the topological differences between these networks. Table 6.4.7 presents the topology information of the analysed networks. With fewer nodes and more links in PWr than Facebook network, the PWr Email network is denser than Facebook network with density of 0.0015 over 0.00042. Nodes in PWr are also highly clustered compared to that in Facebook network as we can see the GCC of PWr is 0.43 which is much higher than 0.1 of Facebook. PWr network has larger diameter than Facebook network although its ASP is smaller. The Flickr network is denser than others with the highest density of 0.0023. The Flickr ASP is 2.329 which means all the nodes in the network are close to each other. The GCC of Flickr network is 0.34, not as high as PWr network. This is because the university email communication is more clustered as users have real world connections. The Twitter network is sparse when compared to other networks.

Network	Nodes	Links	Ave Degree	Ave Shortest Path	Diameter	Clustering Coefficient	Density
Facebook	7,446	23,443	6.297	5.455	15	0.1	0.0008
PWr	6,059	27,640	9.124	4.363	20	0.43	0.0015
Twitter	1,564	2,376	2.554	9.544	33	0.139	0.0019
Flickr	5,949	408,086	137.2	2.329	6	0.34	0.0023

Table 6.11: Network Topology Information

Different network prediction results are observed for different types of networks. Thus, it is necessary to compare the real world network with classic network models like random network model (Section 2.3.2) and regular network model (Section 2.3.1). To do it, we calculated the theoretical GCC and ASP in regular and random networks with the same number of nodes and links as in the real world network using the fomulas stated in Table 6.4.7. The results are presented in Table 6.4.7. For the four networks, their GCC and ASP lies between random network and regular network. The ASPs of real world networks are all very close to the random networks. As for the GCC, the Facebook, Flickr and Twitter networks are closer to the random network while the PWr network is closer to the regular network than other analysed networks. Considering the power law degree distribution, as shown from Fig 6.38 to Fig 6.41, we can conclude that all analysed networks are a combination of small–world and scale–free networks.

The difference in GCC in these networks is caused by the nature of human social networks. For a friendship network like Facebook and Flickr, one's friend may not know all friends of this friend and additionally we tend to have more relationships online than in real world. The clustering coefficient, thus, may not be very large. However, for Email communication network within a large organization, there are always smaller cliques in which people know each other quite well. It could be a department or here in this case, people in the same university. Networks like this are highly clustered and nodes could also be more active within a clique. This leads to a highly clustered dense network that is just like the PWr network.

To gain a better understanding of the network, we generate the overview for the four networks as shown from Fig 6.42 to Fig 6.45. They are created with Gephi [105]. Nodes in the same community are labelled with same colour and the communities are detected using the Louvain method [106], a mod-

Metrics	Random Network	Regular Network
GCC	$\frac{k}{n}$	$\frac{3(k-2)}{4(k-1)}$
ASP	$\frac{\log n}{\log k}$	$\frac{n}{2k}$

Table 6.12: Analytical formulas for CC & ASP in random and regular networks

	Random Network	Facebook	Regular Network
Nodes	7,446	7,446	7,446
Links	23,443	23,443	23,443
GCC	0.000846	0.1	0.6084
ASP	4.845	5.455	591.233
	Random Network	PWr	Regular Network
Nodes	6,059	6,059	6,059
Links	27,640	27,640	27,640
GCC	0.0015	0.43	0.6577
ASP	3.939	4.363	372.907
	Random Network	Flickr	Regular Network
Nodes	5,949	5,949	5,949
Links	408,086	408,086	408,086
GCC	0.023	0.034	0.744
ASP	1.766	2.329	21.712
	Random Network	Twitter	Regular Network
Nodes	1,564	1,564	1,564
Links	2,376	2,376	2,376
GCC	0.0016	0.0139	0.1857
ASP	7.848	9.544	306.186

Table 6.13: Theoretical and Real Network Metrics Facebook & PWr & Flickr & Twitter

ularity based detection methods integrated in Gephi. We observed 28 small communities in PWr email communication network, a lot more than Facebook network in which we find only 12 communities. This explained the reason why PWr has a higher GCC. In Twitter network, we find 16 communities while in Flickr there are only 7 communites been detected. Different dynamics of evolution for different networks, as observed in our results, led to different topology structures.

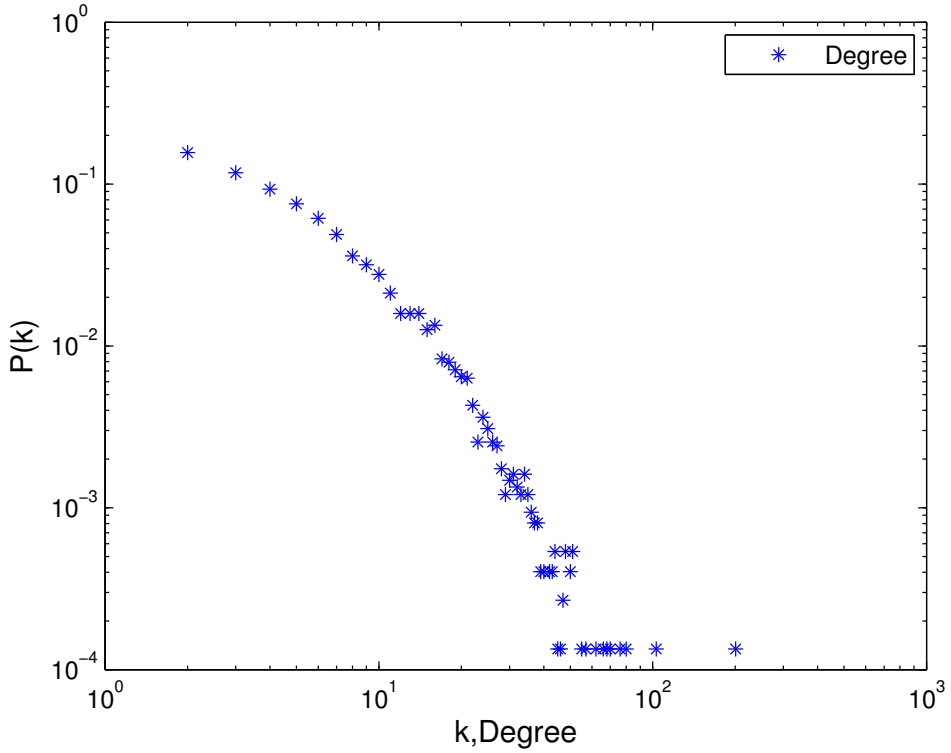


Figure 6.38: Facebook Daily Sliding Window Method Weight

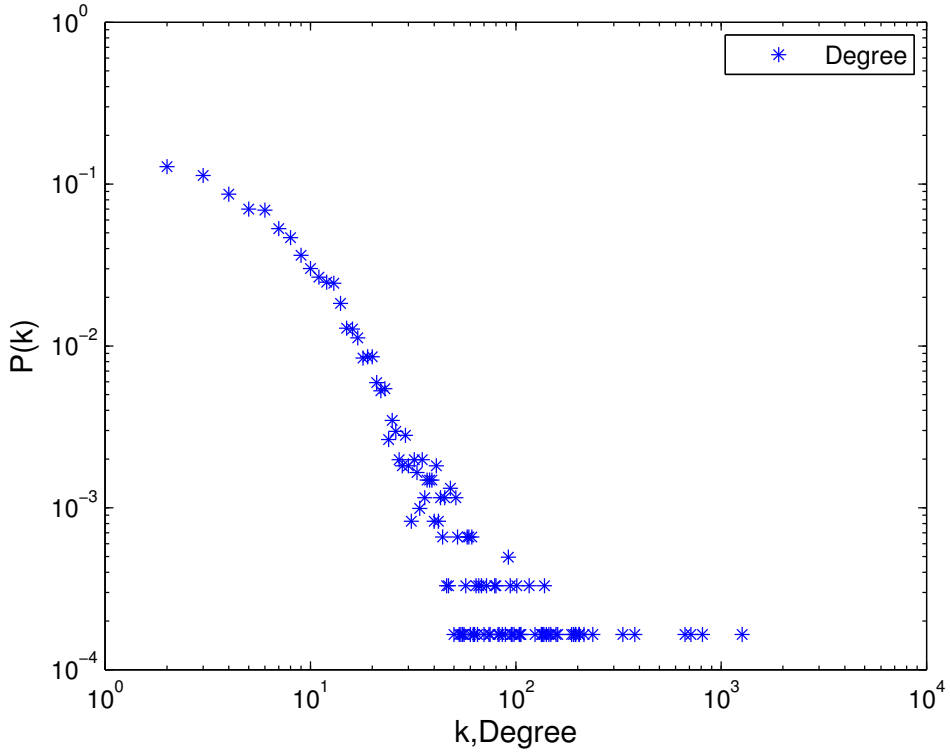


Figure 6.39: PWr Sliding Window Method Weight

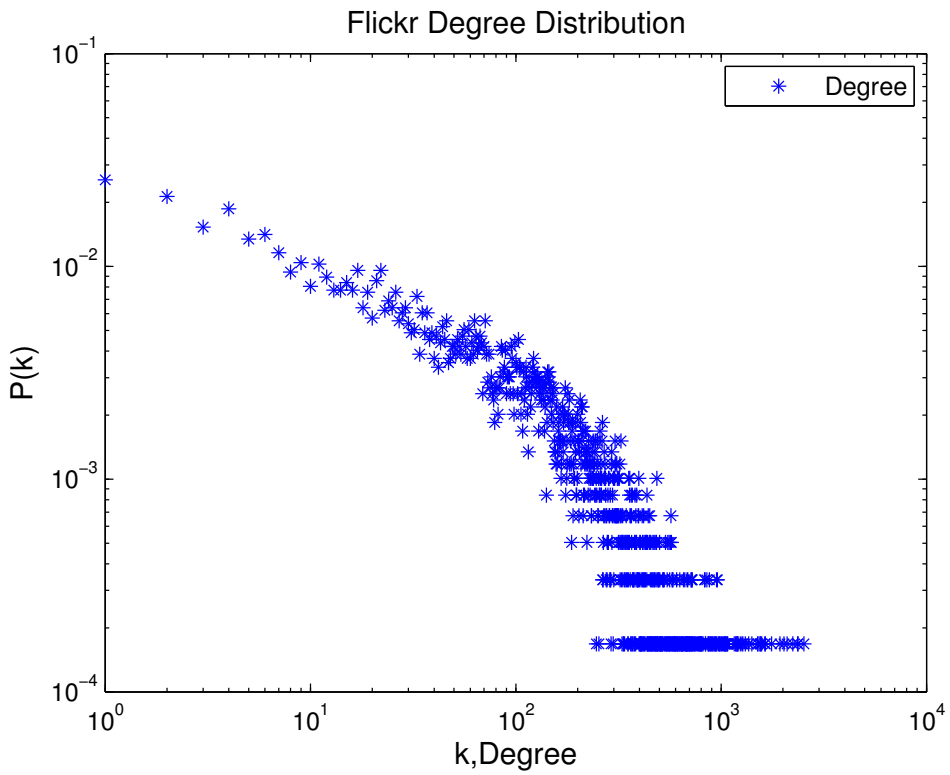


Figure 6.40: Flickr Degree Distribution

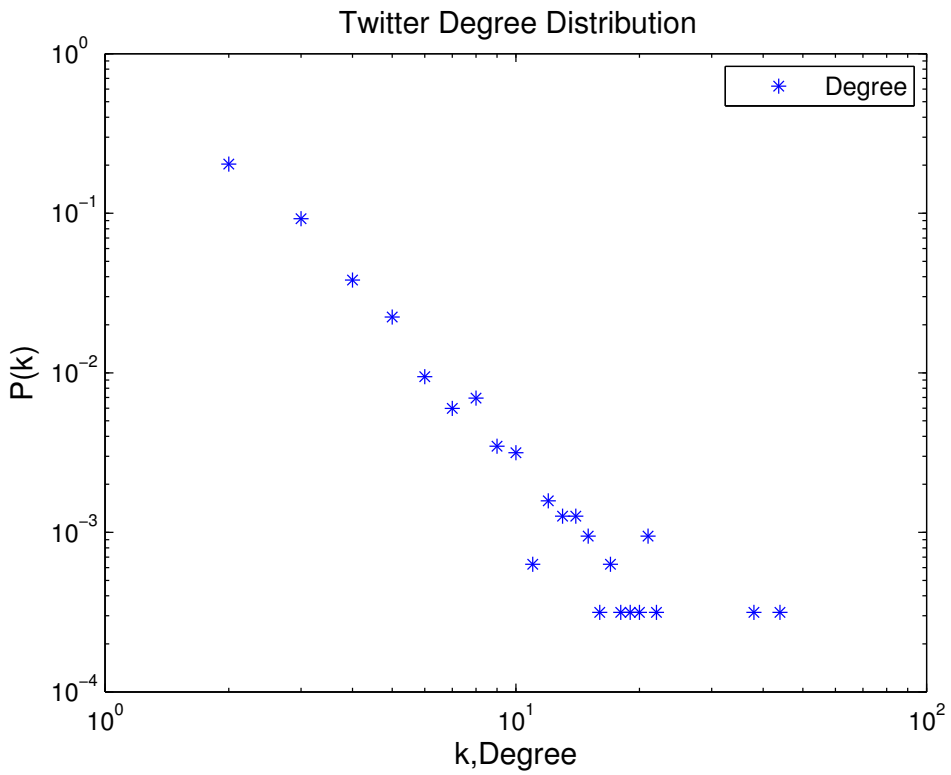


Figure 6.41: Twitter Degree Distribution

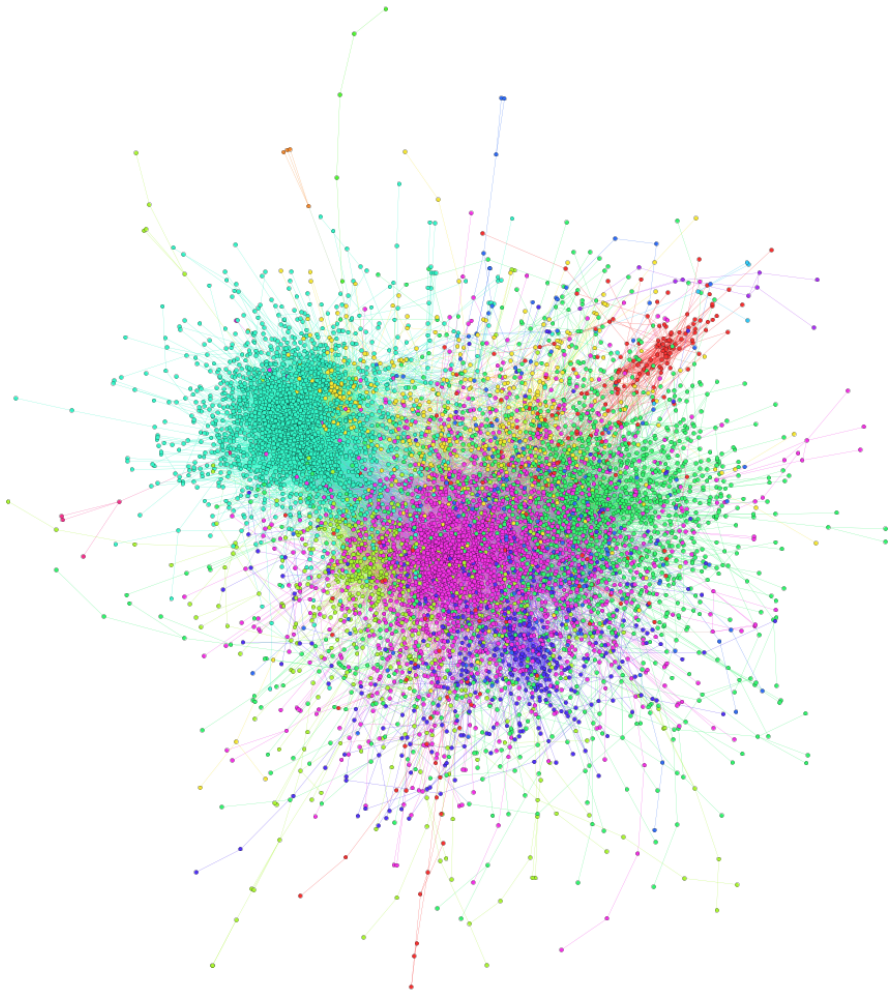


Figure 6.42: Facebook Network Overview (12 Communities)

Figure 6.43: PWr Network Overview (28 Communities)

Figure 6.44: Flickr Network Overview (7 Communities)

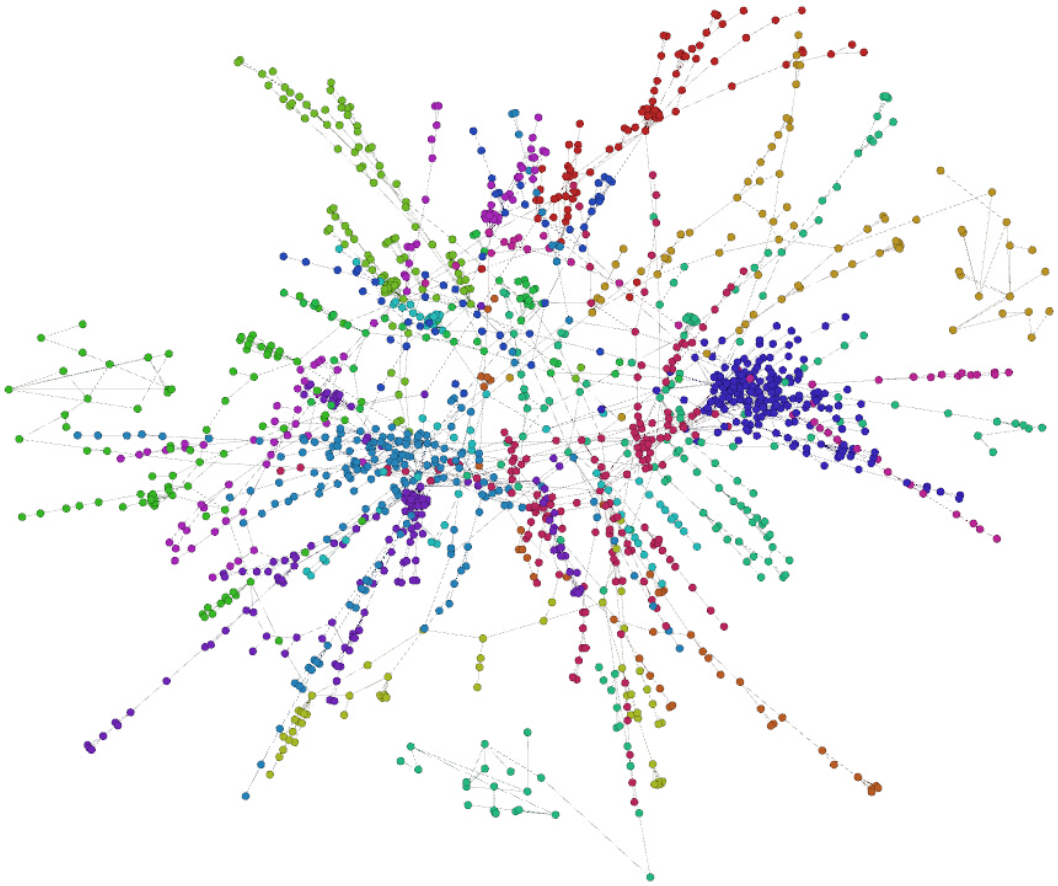


Figure 6.45: Twitter Network Overview (16 Communities)

6.5 Conclusion

In this chapter, we claim that online social networks evolve following certain rules that may change over time. Taking that into account, we introduced new hybrid link prediction model which was tested on four real world online social networks. Selected networks are of different types: (i) friendship network (Facebook), (ii) activity-based network (PWr email network), (iii) tag message network ('@username' in Twitter), and (iv) who follows who network (Flickr). The results of the experiments show that the prediction accuracy of hybrid model is higher than any of the other eight tested methods. Although the model outperforms all other eight selected methods, it still has a limit. As the model is a combination of selected methods, its prediction results heavily relies on the results of selected methods which means it could not predict new links other than the ones predicted by the selected methods. It also explains why the changes in the precision level of the hybrid model always follows the changes in the precision level of other well performing methods as seen in Fig 6.2 - Fig 6.23 which is created using Gephi.

The prediction precision and methods weights results of the four networks are different. For Facebook network, the average prediction precision of hybrid model with growing window scenario is better than that with sliding window scenario. It is contrary to the results for PWr and Flickr network in which the hybrid model sliding window scenario results are much better than that of growing window scenario. This is due to the link timeliness difference of the two types of networks. In Email communication network, links are formed by Email sending between two nodes. This type of links is only valid for a few days and thus accumulate topology information by growing window does not help a lot for link prediction task. This explanation also applies to the Flickr networks. In such photography interested based network, links are formed driven by user

interests and account popularity and activity which are keep changing as time goes by. So the recently formed links are more important than old ones and thus sliding window scenario outperforms growing window scenario. Same for Twitter network, user is less likely to reply to a message that was sent many days ago which means links are valid for a limit period. However, due to the network we used for experiment only contains one week information, the phenomenon of sliding window prediction better than growing window prediction is not reflected in the result. We can obtain the conclusion that for the type of networks with links valid for a long time, the hybrid model with growing window prediction is better than sliding window prediction. However, for networks with links only valid for a short period, sliding window prediction should outperform growing window prediction.

The methods weights results also give a hint that the analysed networks evolved in different ways. The methods weights we obtained are relative stable in Facebook network comparing to PWr network as we can see that the width of different colour stripes in Fig 6.24 - Fig 6.27 did not change as much as that in Fig 6.28 - Fig 6.31 respectively. Similar to the Facebook network, Twitter and Flickr network are also relative stable as shown from Fig 6.32 to Fig 6.37 compare to the PWr network. We can conclude that the networks are evolving in different ways and also lead to different topology structures as shown in Fig 6.42, Fig 6.44, Fig 6.45 and Fig 6.43. Selecting the proper experiment scenario (i.e. Sliding window or Growing window) could help improve the prediction accuracy of our hybrid model.

To summarize, the hybrid model can help improve prediction accuracy. However, as a combination method, it has limitations that the model cannot predict new links that not predicted by the combined methods. For networks with links that valid for short period of time, sliding window perform better while for networks with links can last longer, growing window perform better.

Meanwhile, the weight of different methods reflect how network evolves.

Although the hybrid model outperforms other methods with a significant percentage increase, the absolute prediction precision value still has a big room for improvement. In our experiment, we only applied the eight well-known prediction methods, but there is more other topology information we can use for a better prediction result. As the networks are the real-world social networks, we can focus on more social behaviour related information to improve link prediction. One of the well-studied social behaviour is network community. Thus, in the next chapter we present Community Bridge Boosting prediction method that helps to improve prediction accuracy by utilising the community and bridge information about the network.

Chapter 7

Community Bridge Boosting Prediction Model

The Community Bridge Boosting Prediciton Model (CBBPM) is introduced in this chapter. The model assumptions and experiment design are stated followed by the experiment results. The conclusions are drawn in the last section of this chapter.

7.1 Study Background and Motivation

Research indicates that different parts of networks evolve in different way, e.g. rich get richer concept from BA model (described in Section 2.3.4) or friend of my friend is my friend phenomenon in Common Neighbour model. This might suggest that applying different approaches for different parts of networks may enhance prediction results. Many authors claim that information about communities existing within a social network can help improve the prediction accuracy [107; 108]. For example, the link prediction methods introduced in [109] reflects the prediction precision can be improved by considering community information with different resolutions. Also, researchers in [110]

showed that an enhanced link prediction method that modifies local similarity measures, such as common neighbour and resource allocation, could improve link prediction precision.

Motivated by this, we decided to look at the communities and bridge nodes linking those communities. Looking at the groups within a network, nodes can be categorised into two types. Type one represents a node for which both this node and all it connections belong to one community and we call it **within community node**. The other type is the node that is connected to different communities and we call it **inter community node**. The novelty of this study is that we are aiming at improving the link prediction accuracy by providing a different treatment for these two types of nodes in one model. We think that nodes connected to more than one community are more likely to form new links as they have more diverse relationships that nodes mainly connected to one community. Proposed below method is based on this assumption.

In this chapter, we will introduce our Community Bridge Boosting Prediction Model and the results of the experiments that we run for the introduced model with different settings. The results will be analysed and summarized in the end.

7.2 Community Bridge Boosting Prediction Model

In this section, we propose the Community Bridge Boosting Prediction Model. This model is a structure based method as introduced in Section 2.4. It also takes the network community information into consideration. Our approach, in contrary to how other authors enhance the similarity score between nodes within the same community in [110], focuses on the nodes that are connected

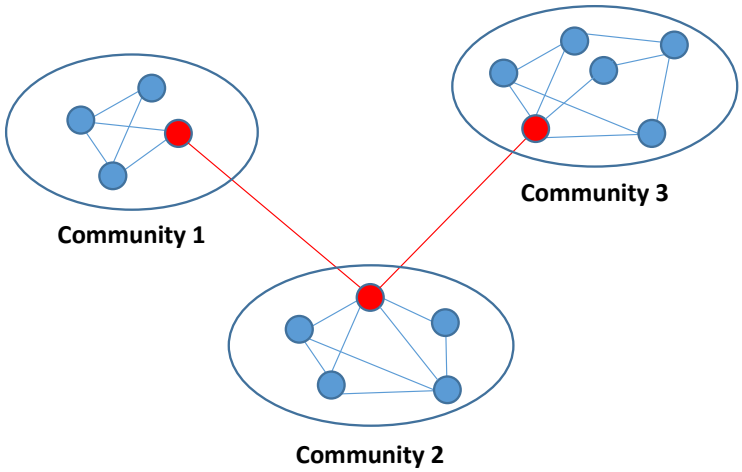


Figure 7.1: Community Bridge Nodes and Links Example (both shown in red)

to the multi-communities. We define the links with two end points located in different communities as bridge link and the two nodes of the bridge link as bridge nodes. Figure 7.1 is an illustration of the concept in which the red nodes and links are the bridge nodes and bridge links.

The bridge nodes are the only connections between communities so that they can play an important role in forming new links. In addition, those are the nodes with high node position [111]. However, in our model, we believe some of the bridge nodes are not playing more important role than the other bridge nodes from the perspective of link prediction task. We think there are two types of bridge nodes that are less important:

- **Type I.** Bridge nodes that are not widely connected in the network so that they have very low degree. Their influence on new link formation is limited due to their limited connection resources.

- **Type II.** For bridge nodes with the majority of its links located in one community while the proportion of its connections to other communities are very small. The importance of such bridge nodes on new link formation is limited because their connection resources are too concentrated in one community.

To reflect the importance of bridge nodes that do not belong to these two types, our proposed model doubles (increase BNSS by 100%) the Bridge Nodes Similarity Score (BNSS). The Type I bridge nodes can be eliminated by setting a degree threshold. For Type II bridge nodes, we proposed a novel rate, the Max Community Dominant Rate (MCDR) to quantify the bridge node link proportion. For each node x, we count the number of links that it is connected to different communities as shown in Equation 7.1

$$\begin{aligned} LinkNum(x)_{All\_Com} = LinkNum(x)_{Com1} + LinkNum(x)_{Com2} \\ +, \ldots, +LinkNum(x)_{ComM} \end{aligned} \tag{7.1}$$

where $LinkNum(x)_{All-Com}$ equals to the total degree of node x and $LinkNum(x)_{ComM}$ stands for the number of links to which x is connected and that come from community M. The MCDR(x) is then defined as:

$$MCDR(x) = \frac{MAX(LinkNum(x)_{Com1}, \ldots, LinkNum(x)_{ComM})}{LinkNum(x)_{All-Com}} \tag{7.2}$$

MCDR is a number between 0 and 1. A node with high MCDR means that the majority of the relationships of this node is within one community. Small MCDR indicates that there are no dominant communities that the node is connected to. We designed a filter to select the bridge nodes that are not classified as the two types of less important bridge nodes. The rules are as follows:

1. Select the bridge nodes with degree larger than the network average degree so that the nodes we selected are well connected in the network.

2. Select the bridge nodes with MCDR smaller than a chosen value R to

guarantee that for the bridge nodes we selected there are (1-R) of its connections connected to other communities.

The resulting Bridge Node Similarity Scores are boosted in the model for link prediction.

7.3 Experiment Design

The whole experiment is guided by the methodology introduced in Figure 3.4. Experiment are following these steps:

1. Calculate the prediction precision result for each network using all the selected methods as the benchmark.

2. Detect the communities in the train network for each dataset. Based on the community result, find all the bridge nodes.

3. Apply the filters to filter out the bridge nodes that we would like to boost. In the filter, we select R as 0.7 0.8 and 0.9. So that the bridge nodes selected have certain amount of friends from other communities. We believe it is more likely for this kind of nodes to form new links.

4. Boost the similarity scores for the selected bridge nodes by doubling them. Use the boosted similarity score to predict new links and calculate the precision.

7.3.1 Datasets

The experiment of this model test is performed on six selected networks includes Enron Email Network, Facebook Wall post Network, Flickr Network, PWr Email Network, UC Irvine Message Network and YouTube Network. The

network details and how they are processed have been introduced in Section 4.2.3.

7.3.2 Selected Methods

The base prediction methods selected in this experiment are introduced in Section 2.4, includes:

- Common Neighbours (CN),
- Jaccard's Coefficient Index (JI),
- Preferential Attachment (PA),
- Adamic/Adar Index (AA),
- Resource Allocation (RA)
- Cosine Similarity (Cos),
- Sorensen Index (Sor),
- Katz method (Katz).

7.3.3 Community Detection

One of the key parts of the Community Bridge Boosting Prediction Method is the community detection. In this study, we adopt the greedy modularity optimization approach to detect communities. The principle of the modularity is that a good community has many internal links but is mostly isolated from the rest of the network [110]. The modularity of a partition is defined as follows: Let m be the number of links in the network, A_{ij} be the number of links between nodes i and j, k_i be the degree of node i, and $\delta(i, j)$ be 0 if i

and j are in different parts of the partition and 1 if they are in the same part. Then the modularity Q of a partition is:

$$Q = \frac{1}{2m} \sum_{i,j} \Big[A_(i,j) - \frac{k_i k_j}{2m} \Big] \delta(i,j) \tag{7.3}$$

In this thesis, we selected a well-known method [112], Louvain method, for the greedy modularity optimization [106].

7.3.4 Prediction Accuracy Measure

In this experiment, the prediction performance is measured using precision (Section 3.5.1). Precision is numbers between 0 and 1. A higher precision means a better prediction accuracy. As we are more interested to investigate the effect of model variable, MCDR (the R value in Step 3), on link prediction accuracy, we only predict the same number of links as the number of new formed links to see how many of them are correctly predicted. In this setting, the precision and recall are same and this is why only precision is used to measure the model performance.

7.4 Experiment Result

For each network, we predicted the new links using different prediction methods with four different experiment settings. (1) Original prediction method prediction. (2) Boost with $R < 0.7$ prediction. (3) Boost with $R < 0.8$ prediction. (4) Boost with $R < 0.9$ prediction.

From Tables 7.1 to Table YY, numbers in bold indicate the best precision results in a given column. The underlined number means it is the best performing method among all the results for a given dataset. Below, results for each of the selected dataset are discussed. In addition, the reasons for obtained

	CN	JA	PA	AA	RA	Cos	Soren	Katz
Original	0.025	0.0000	**0.011**	0.022	0.014	0.000	0.000	**0.026**
Boost (R<0.7)	0.024	0.0000	**0.011**	0.022	0.014	0.000	0.000	0.025
Boost (R<0.8)	0.024	0.0000	0.010	0.021	**0.016**	0.001	**0.001**	**0.026**
Boost (R<0.9)	**0.028**	0.0000	0.010	**0.024**	0.014	**0.002**	**0.001**	0.025

Table 7.1: Enron Network CBBPM Prediction Precision Result

prediction accuracies are analysed and presented.

7.4.1 Enron Network

Table 7.1 shows the prediction result for Enron network for original methods as benchmarks as well as the proposed Community Bridge Boosting method using different MCDRs. As we can see in the Table 7.1, the proposed method always performs better or equally as good as the original method. For Cosine Index prediction and Soren Index prediction, the original method did not predict any new links correctly with a precision of 0 while the boosted method, for $R < 0.8$ and $R < 0.9$, is capable to predict new links with precision of 0.002 and 0.001 respectively.

In the original prediction methods, the best performed method is Katz with precision of 0.026. However, among all the prediction result, the common neighbour with $R < 0.9$ boost gives the best prediction result with prediction precision of 0.028. It improved 11% against the original common neighbour method. Comparing original Katz, it still gives an improvement of 9%.

7.4.2 Facebook Wall Post Network

Prediction results for Facebook network are summarized in Table 7.2. In the Facebook wall post prediction experiment, among all the prediction method,

the best performed method, with precision of 0.036, is Adamic / Adar Index using $R < 0.9$ boost. It gives an improvement of 5% in comparison to the original method. Meanwhile, five out of eight $R < 0.9$ boosted methods give the best prediction result against original method and other boosting settings.

	CN	JA	PA	AA	RA	Cos	Soren	Katz
Original	**0.033**	0.014	**0.002**	0.034	**0.034**	0.011	0.014	0.026
Boost (R<0.7)	0.025	0.016	0.000	0.028	0.026	0.012	0.015	0.026
Boost (R<0.8)	0.027	0.018	0.000	0.032	0.027	0.013	0.017	0.026
Boost (R<0.9)	0.026	**0.019**	0.000	**0.036**	0.024	**0.020**	**0.022**	**0.027**

Table 7.2: Facebook Network CBBPM Prediction Precision Result

7.4.3 Flickr Network

The experiment results from Table 7.3 with Flickr network is different from the first two experiments. We do not observe improvement of our proposed method against original network. Resource Allocation method gives the best prediction result, precision 0.0345, together with $R < 0.7$ boost method. Comparing with most of the methods, the community bridge boosting prediction method delivered same precision results.

	CN	JA	PA	AA	RA	Cos	Soren	Katz
Original	**0.027**	**0.028**	**0.023**	**0.028**	**0.035**	**0.027**	**0.028**	**0.027**
Boost (R<0.7)	**0.027**	**0.028**	0.022	**0.028**	**0.035**	**0.027**	**0.028**	**0.027**
Boost (R<0.8)	**0.027**	**0.028**	0.022	**0.028**	0.034	**0.027**	**0.028**	**0.027**
Boost (R<0.9)	**0.027**	**0.028**	0.022	**0.028**	0.034	**0.027**	**0.028**	**0.027**

Table 7.3: Flickr Network CBBPM Prediction Precision Result

7.4.4 PWr Email Network

Table 7.4 presents the PWr Email network prediction results. Our community bridge boosting prediction method could always deliver better or same prediction result. The best original prediction method is Adamic / Adar Index with precision of 0.011. The $R < 0.7$ boost method push the precision result up by 16% to 0.013 which over perform all the other methods.

	CN	JA	PA	AA	RA	Cos	Soren	Katz
Original	**0.006**	0.001	0.000	0.011	0.007	0.000	0.001	**0.006**
Boost (R<0.7)	**0.006**	0.005	0.000	**0.013**	**0.009**	0.005	**0.007**	**0.006**
Boost (R<0.8)	**0.006**	**0.006**	0.000	0.009	**0.009**	**0.006**	0.006	**0.006**
Boost (R<0.9)	**0.006**	0.004	0.000	0.008	**0.009**	0.004	0.004	0.004

Table 7.4: PWr Email Network CBBPM Prediction Precision Result

7.4.5 UC Irvine Message Network

Our proposed community bridge boosting prediction method is capable to provide a prediction precision improvement in UC Irvine message network experiment as shown in Table 7.5. The best performed method among the original methods is Preferential Attachment with precision of 0.022. The $R < 0.7$ boost method makes the precision up by 17% better to 0.025 than the original method.

7.4.6 YouTube Network

As shown in Table 7.6, the community bridge boosting prediction method does not improve the prediction precision. The prediction result of $R < 0.7$ boost method is same as the original method for Common Neighbour and Katz method. The best prediction method for YouTube network is original Katz

	CN	JA	PA	AA	RA	Cos	Soren	Katz
Original	**0.014**	**0.001**	0.022	0.015	0.013	0.001	0.001	0.014
Boost (R<0.7)	0.013	**0.001**	**0.025**	0.013	**0.016**	**0.002**	**0.002**	0.013
Boost (R<0.8)	**0.014**	0.001	0.022	**0.016**	0.013	**0.002**	**0.002**	**0.015**
Boost (R<0.9)	**0.014**	**0.001**	0.022	0.015	0.013	**0.002**	**0.002**	0.014

Table 7.5: UC Irvine Network CBBPM Prediction Precision Result

together with $R < 0.7$ boost method.

	CN	JA	PA	AA	RA	Cos	Soren	Katz
Original	**0.041**	0.000	**0.026**	**0.039**	**0.029**	0.000	0.000	**0.042**
Boost (R<0.7)	**0.041**	0.000	0.024	**0.039**	0.027	0.000	0.000	**0.042**
Boost (R<0.8)	0.037	0.000	0.024	0.036	0.027	0.000	0.000	0.038
Boost (R<0.9)	0.034	0.000	0.021	0.032	0.023	0.000	0.000	0.035

Table 7.6: YouTube Network CBBPM Prediction Precision Result

7.5 Conclusion

Out of the six CBBPM experiments we performed on different networks, four experiments provide improvement of prediction precision and the other two experiments on YouTube network and Flickr network show no enhancement. For these two networks, the CBBPM only gives equal prediction precision as the original prediction methods in the best case (YouTube network Katz method with Boost ($R < 0.7$) and Flickr network RA method with Boost ($R < 0.7$)). An explanation for this phenomenon is the different nature of networks. YouTube and Flickr network are both subscribe based networks that based on user interest. Such networks are more relevant to personal preference which is not likely to change dramatically. Once the interest based

Networks	Boost$_{R<0.7}$ Nodes Num	Boost$_{R<0.8}$ Nodes Num	Boost$_{R<0.9}$ Nodes Num	Total Nodes Num
UC Irven	644	947	1086	1666
Enron	1283	1688	2020	5738
Flickr	95	147	201	4200
Facebook	1825	2334	2587	5784
Pwr	899	1256	1525	6335
Youtube	620	946	1337	6000

Table 7.7: CBBPM Nodes Boosting Number

communities are formed in this kind of networks, the influence of bridge nodes on new link formation is limited. Because bridge node users have different interests which led them connected to different communities and this may not affect other users' interests. Thus boosting bridge nodes in such interested driven networks cannot improve prediction accuracy and this is also observed in our experiments. Table 7.7 states the number of bridge nodes we selected for each network with different R value. We can see for both Flickr and YouTube network, the number of bridge nodes selected are smaller than other networks except UC Irvine, which is a small network contains only 1,666 nodes compare to others. This also proved our explanation.

Among the four experiments in which CBBPM gives a prediction precision enhancement, the best performed boost prediction result is obtained with two settings: $R < 0.7$ for UC Irvine message and PWr Email networks. $R < 0.9$ for Enron Email and Facebook Wall Post networks. So there is no optimized R value that works for all networks. The best R value selection could be a potential future study of this model.

To summarize, our CBBPM is capable to predict links with the same or better precision than traditional prediction methods. The R value is important to the prediction accuracy which needs further investigation about the best R value selection.

Chapter 8

Conclusion and Future Work

In this last chapter, we review the research hypotheses and research questions. The findings and conclusions are then summarized to show haw the hypotheses were verified. The last part introduces potential future works of my study.

8.1 Conclusion

This thesis analysed online social network link prediction problem. We believe a dynamic prediction approach that treats nodes or links differently depending on the networks components characteristics and structure could be a good approach to improve the prediction performance. Our research is guided by two hypotheses:

- Hypothesis 1. The performance of structure based network prediction methods and the characteristics of the networks are correlated.
- Hypothesis 2. As networks are dynamic, the performance of prediction can be improved by providing different treatment to different nodes and links.

To verify Hypothesis 1, we studied the Pearson correlation between the performance of ten network structure based link prediction methods and six selected network metrics. We found very strong positive correlation between gini coefficient and preferential attachment prediction accuracy with a correlation coefficient of 0.94. Cosine similarity and sørensen index prediction accuracy is also highly positively correlated to global clustering coefficient with Pearson correlation of 0.8 and 0.81 respectively. Meanwhile, preferential attachment is also negative correlated to network diameter and average shortest path with coefficient of -0.77 and -0.79 (Table 5.5). Hypothesis 1 is then verified. But we need to point out that a high correlation between link prediction method performance and network metrics does not mean the method has better prediction accuracy than other methods. Another finding from this study is the classification of networks into two subsets: (i) prediction friendly networks and (ii) prediction unfriendly networks. Prediction friendly network refers to networks that could be better predicted (with AUC over 0.8) by most of the prediction methods while for prediction unfriendly network is the other way around with most prediction method AUC result under 0.8. We found that networks with the structural profile similar to smallworld network are easier to predict than networks similar to random structures. This work has been published in [77].

We then proposed two models to verify Hypothesis 2. The hybrid model is proposed and based on the assumption that network links are formed following some patterns. The model supports two predicting scenarios: sliding window and growing window. Both scenarios were tested with four networks. On average, the hybrid model outperforms other prediction methods in all of the four networks. However, there is a limit for our hybrid model. As a model combines several methods, the prediction accuracy heavily relies on the selected methods which means it could not predict new links other than the ones predicted by

the selected methods. We can conclude that network links are formed following different rules and these rules and their co-occurrence cause high complexity and dynamic of online social network evolution. This can also be observed in Fig 6.24 to Fig 6.37 as the weight for each combined methods keep changing. Also, prediction performances of the two model scenarios (sliding v. growing window) are also network dependent. For networks with links that are valid for a short period of time, like Flickr and PWr network in our experiment, the performance of sliding window prediction is better than growing window prediction. If links are valid for a long time in the network, then growing windows prediction supposed to perform better as shown in Facebook friend network experiment. In [113] we published the results and findings of this study.

Different from the hybrid model that focused on the links, we proposed Community Bridge Boosting Prediction Model which aimed at providing different treatments for nodes. This model uses community information for link prediction test. Experiment results showed that this model is capable to improve the prediction accuracy for networks that are likely to have inter-community interactions like the PWr Email network in our experiments. The key variable in this model is the Max Community Dominant Rate which is used to select bridge nodes for boosting. We selected the rate as $0.7, 0.8$ and 0.9 in our experiment and prediction accuracy varies over different rate values. We did not further analyse the best rate selection as it is not the main purpose of this study.

To sum up, the two hypotheses stated in Chapter 1 are verified by three groups of experiments. The two link prediction models can help improving network link prediction accuracy and also inspire further research on dynamic social network evolvement from the perspective of link prediction.

8.2 Future Work

There are several potential topics inspired by this thesis that can be extended. In Prediction Accuracy and Network Metrics study, we found the prediction friendly networks and prediction unfriendly networks. The reason behind this was not fully investigated in this thesis. Finding the threshold to classify the two types of the networks could be next research to do which may potentially benefit link prediction research as well. Another interesting further research area is the evolution of network dynamics. In the hybrid model experiment, we observed the network evolution from the weights of the ten selected prediction methods used in the combination. A more systematic study on dynamics of network evolution from the angel of link prediction would be also a good direction for future works. The way to best select the optimal Max Community Dominant Rate in our Community Bridge Boosting Prediction Model is also a valid research topic. In this thesis, we proposed two different prediction models, it would also be a worth looking topic to combine the two models to predict new links as both time-scale network evolution information and network community information could be used for link prediction.

References

[1] N. Biggs, E. K. Lloyd, and R. J. Wilson. *Graph Theory, 1736-1936*. Clarendon Press, New York, NY, USA, 1986. 1

[2] P. Erdős and A. Rényi. On random graphs. I. *Publ. Math. Debrecen*, 6:290–297, 1959. 1, 19

[3] P. Erdös and A Rényi. On the evolution of random graphs. In *PUBLICATION OF THE MATHEMATICAL INSTITUTE OF THE HUNGARIAN ACADEMY OF SCIENCES*, pages 17–61, 1960. 1, 19

[4] Jeffrey Travers, Stanley Milgram, Jeffrey Travers, and Stanley Milgram. An experimental study of the small world problem. *Sociometry*, 32:425–443, 1969. 2

[5] Stanley Milgram. The Small World Problem. *Psychology Today*, 2:60–67, 1967. 2, 17

[6] D. J. Watts and S. H. Strogatz. Collective dynamics of'small-world'networks. *Nature*, 393(6684):409–10, 1998. 2, 16, 17, 23, 68

[7] A. L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999. 2

[8] Jing-Dong J. Han, Nicolas Bertin, Tong Hao, Debra S. Goldberg, Gabriel F. Berriz, Lan V. Zhang, Denis Dupuy, Albertha J. M. Wal-

hout, Michael E. Cusick, Frederick P. Roth, and Marc Vidal. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, 430(6995):88–93, July 2004. 2

[9] Tong Hao, Wei Peng, Qian Wang, Bin Wang, and Jinsheng Sun. Reconstruction and application of proteinprotein interaction network. *International Journal of Molecular Sciences*, 17(6):907, 2016. 2

[10] Fan Zhang, Won-min Song, SiDe Li, Vashisht Ajay, Francesca Aguilo, Anindya Bagchi, James A. Wohlschlegel, Bin Zhang, and Martin Walsh. Abstract pr06: The enhancer landscape involves a core noncoding rna protein interaction network for c-myc expression. *Cancer Research*, 76(6 Supplement):PR06–PR06, 2016. 2

[11] Arndt Grossmann, Nouhad Benlasfer, Petra Birth, Anna Hegele, Franziska Wachsmuth, Luise Apelt, and Ulrich Stelzl. Phospho-tyrosine dependent protein–protein interaction network. *Molecular Systems Biology*, 11(3), 2015. 2

[12] Mason S. P. Barabasi A.-L. Oltvai Z. N. Jeong, H. Lethality and centrality in protein networks. *Nature*, 411:41–42, 2001. 2, 11, 12

[13] A. D. King, N. Pržulj, and I. Jurisica. Protein complex prediction via cost-based clustering. *Bioinformatics*, 20(17):3013–3020, November 2004. 2, 11

[14] Marie Chevallier, Meziane Aite, Jeanne Got, Guillaume Collet, Nicolas Loira, Maria-Paz Cortes, Clémence Frioux, Julie Laniau, Camille Trottier, Alejandro Maas, and Anne Siegel. Handling the heterogeneity of genomic and metabolic networks data within flexible workflows with the PADMet toolbox. In *Jobim 2016 : 17ème Journées Ouvertes en Biologie, Informatique et Mathématiques*, Lyon, France, June 2016. 2

[15] Hojung Nam, Miguel Campodonico, Aarash Bordbar, Daniel R. Hyduke, Sangwoo Kim, Daniel C. Zielinski, and Bernhard O. Palsson. A systems approach to predict oncometabolites via context-specific genome-scale metabolic networks. *PLOS Computational Biology*, 10(9):1–13, 09 2014. 2

[16] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.*, 86(14):3200–3203, 2001. 2

[17] Yang Wang, Deepayan Chakrabarti, Chenxi Wang, and Christos Faloutsos. Epidemic spreading in real networks: An eigenvalue viewpoint. In *In SRDS*, pages 25–34, 2003. 2

[18] Deepayan Chakrabarti, Yang Wang, Chenxi Wang, Jurij Leskovec, and Christos Faloutsos. Epidemic thresholds in real networks. *ACM Trans. Inf. Syst. Secur.*, 10(4):1:1–1:26, January 2008. 2

[19] Nicholas A. Christakis and James H. Fowler. The spread of obesity in a large social network over 32 years. *New England Journal of Medicine*, 357(4):370–379, 2007. 2, 11

[20] Zan Huang, Xin Li, and Hsinchun Chen. Link prediction approach to collaborative filtering. In *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, JCDL '05, pages 141–142, New York, NY, USA, 2005. ACM. 2, 83

[21] Ferenc Molnar. Link Prediction Analysis in the Wikipedia Collaboration Graph, 2011. 2, 12, 83

[22] Fahimeh Ghasemian, Kamran Zamanifar, Nasser Ghasem-Aqaee, and Noshir Contractor. Toward a better scientific collaboration success pre-

diction model through the feature space expansion. *Scientometrics*, 108(2):777–801, 2016. 2

[23] Young-Ho Eom and Hang-Hyun Jo. Generalized friendship paradox in complex networks. *CoRR*, abs/1401.1458, 2014. 2

[24] Chris G. Antonopoulos, Shambhavi Srivastava, Sandro E. de S. Pinto, and Murilo S. Baptista. Do brain networks evolve by maximizing their information flow capacity? *PLOS Computational Biology*, 11(8):1–29, 08 2015. 2

[25] Lada Adamic and Eytan Adar. Friends and neighbors on the web. *Social Networks*, 25:211–230, 2001. 2, 11, 31

[26] W. Cukierski, B. Hamner, and Bo Yang. Graph-based features for supervised link prediction. In *Neural Networks (IJCNN), The 2011 International Joint Conference on*, pages 1237–1244, 31 2011-Aug. 5. 2, 30

[27] M. Fire, L. Tenenboim, O. Lesser, R. Puzis, L. Rokach, and Y. Elovici. Link prediction in social networks using computationally efficient topological features. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 73–80, Oct. 2

[28] K. Juszczyszyn, K. Musial, and M. Budka. Link prediction based on subgraph evolution in dynamic social networks. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 27–34, 2011. 2

[29] David Liben-Nowell and Jon Kleinberg. The link prediction problem for social networks. In *Proceedings of the twelfth international conference on Information and knowledge management*, CIKM '03, pages 556–559, New York, NY, USA, 2003. ACM. 3, 5, 7, 8, 11, 28, 30, 31, 33, 35, 85

[30] Zhepeng (Lionel) Li, Xiao Fang, and Olivia R. Liu Sheng. A survey of link recommendation for social networks: Methods, theoretical foundations, and future research directions. *CoRR*, abs/1511.01868, 2015. 3

[31] L. Lü and T. Zhou. Link prediction in complex networks: A survey. *Physica A Statistical Mechanics and its Applications*, 390:1150–1170, March 2011. 3, 7, 8, 11, 30, 33, 34, 35, 53, 55, 67, 85

[32] Andrew Chen-Brian Tran Ole J. Mengshoel, Raj Desai. Will we connect again? machine learning for link prediction in mobile social networks. 2013. 3, 5

[33] Zhen Liu, Qian-Ming Zhang, Linyuan L, and Tao Zhou. Link prediction in complex networks: A local nave bayes model. *EPL (Europhysics Letters)*, 96(4):48007, 2011. 3

[34] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424(45):175 – 308, 2006. 5

[35] M. Budka, K. Juszczyszyn, K. Musial, and A. Musial. Molecular model of dynamic social network based on e-mail communication. *Social Network Analysis and Mining*, 2013. 5

[36] Guido Caldarelli, Alessandro Chessa, Irene Crimaldi, and Fabio Pam-

molli. The Evolution of Complex Networks: A New Framework. March 2012. 5

[37] Mohammad Al Hasan, Vineet Chaoji, Saeed Salem, and Mohammed Zaki. Link prediction using supervised learning. In *In Proc. of SDM 06 workshop on Link Analysis, Counterterrorism and Security*, 2006. 5

[38] Xue-Wen Chen and Mei Liu. Prediction of proteinprotein interactions using random decision forest framework. *Bioinformatics*, 21(24):4394–4400, 2005. 5, 66

[39] Patrick Aloy and Robert B. Russell. Interprets: protein interaction prediction through tertiary structure. *Bioinformatics*, 19(1):161–162, 2003. 5, 66

[40] Jiajun Bu Xin Wang Yue Wu Chun Chen Zhi Yu, Can Wang. Friend recommendation with content spread enhancement in social networks. *Inf. Sci.*, 309(C):102–118, July 2015. 6

[41] Andrew Cook, Henk A.P. Blom, Fabrizio Lillo, Rosario Nunzio Mantegna, Salvatore Miccich, Damin Rivas, Rafael Vzquez, and Massimiliano Zanin. Applying complexity science to air traffic management. *Journal of Air Transport Management*, 42:149 – 158, 2015. 6

[42] M. E. J. Newman. The Structure and Function of Complex Networks. *SIAM Review*, 45(2):167–256, 2003. 7, 11, 68

[43] James Ladyman, James Lambert, and Karoline Wiesner. What is a complex system? *European Journal for Philosophy of Science*, 3(1):33–67, 2013. 11

[44] D.P. Almond, C.J. Budd, M.A. Freitag, G.W. Hunt, N.J. McCullen, and N.D. Smith. The origin of power-law emergent scaling in large bi-

nary networks. *Physica A: Statistical Mechanics and its Applications*, 392(4):1004 – 1027, 2013. 11

[45] James L. Goodson. The vertebrate social behavior network: Evolutionary themes and variations. *Horm Behavious*, pages 11–22, 2005. 11

[46] Swami Iyer and Timothy Killingback. Evolution of cooperation in social dilemmas on complex networks. *PLOS Computational Biology*, 12(2):1–25, 02 2016. 11

[47] Francisco A. Rodrigues, Thomas K. DM. Peron, Peng Ji, and Jrgen Kurths. The kuramoto model in complex networks. *Physics Reports*, 610:1 – 98, 2016. The Kuramoto model in complex networks. 11

[48] Rashid V Williams-Garcia, John M Beggs, and Gerardo Ortiz. Unveiling causal activity of complex networks. *arXiv preprint arXiv:1603.05659*, 2016. 11

[49] Linyuan L, Duanbing Chen, Xiao-Long Ren, Qian-Ming Zhang, Yi-Cheng Zhang, and Tao Zhou. Vital nodes identification in complex networks. *Physics Reports*, 650:1 – 63, 2016. Vital nodes identification in complex networks. 11

[50] S. Boccaletti, J.A. Almendral, S. Guan, I. Leyva, Z. Liu, I. Sendia-Nadal, Z. Wang, and Y. Zou. Explosive transitions in complex networks structure and dynamics: Percolation and synchronization. *Physics Reports*, 660:1 – 94, 2016. Explosive transitions in complex networks structure and dynamics: Percolation and synchronization. 11

[51] Amedeo Caflisch Francesco Rao. The protein folding network. *Journal of Molecular Biology*, 342(1):299 – 306, 2004. 11

[52] Rachel A. Hillmer and Fumiaki Katagiri. Toward predictive modeling of large and complex biological signaling networks. *Physiological and Molecular Plant Pathology*, 95:77 – 83, 2016. The U.S.-Japan Scientific Seminar: Molecular Contact Points in Host-Pathogen Co-evolution. 11

[53] Glen Jeh and Jennifer Widom. Simrank: A measure of structural-context similarity. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 538–543, New York, NY, USA, 2002. ACM. 11

[54] Michael L. Black. The world wide web as complex data set: Expanding the digital humanities into the twentieth century and beyond through internet research. *International Journal of Humanities and Arts Computing*, 10(1):95–109, 2016. 11

[55] Robert W. Taylor, Eric J. Fritsch, and John Liederbach. *Digital Crime and Digital Terrorism*. Prentice Hall Press, Upper Saddle River, NJ, USA, 3rd edition, 2014. 12

[56] Paul A. C. Duijn and Peter P. H. M. Klerks. *Social Network Analysis Applied to Criminal Networks: Recent Developments in Dutch Law Enforcement*, pages 121–159. Springer International Publishing, Cham, 2014. 13

[57] Daniel M. Schwartz and Tony (D.A.) Rouselle. Using social network analysis to target criminal networks. *Trends in Organized Crime*, 12(2):188–207, 2009. 13, 17

[58] Mark D. Leiserson, Fabio Vandin, Hsin-Ta T. Wu, Jason R. Dobson, Jonathan V. Eldridge, Jacob L. Thomas, Alexandra Papoutsaki, Younhun Kim, Beifang Niu, Michael McLellan, Michael S. Lawrence, Abel

Gonzalez-Perez, David Tamborero, Yuwei Cheng, Gregory A. Ryslik, Nuria Lopez-Bigas, Gad Getz, Li Ding, and Benjamin J. Raphael. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature genetics*, 47(2):106–114, February 2015. 13

[59] Mark Newman. *Networks: An Introduction*. Oxford University Press, Inc., New York, NY, USA, 2010. 14, 15, 19, 20, 22, 24, 68, 69, 70

[60] John P. Scott. *Social Network Analysis: A Handbook*. SAGE Publications, January 2000. 17

[61] Qian-Ming Zhang, Linyuan L, Wen-Qiang Wang, Yu-Xiao, and Tao Zhou. Potential theory for directed networks. *PLOS ONE*, 8(2):1–8, 02 2013. 17

[62] Alex Arenas, Antonio Cabrales, Albert Daz-guilera, Roger Guimer, and Fernando Vega-redondo. Search and congestion in complex networks. In *Proceedings of the XVIII Sitges Conference on Statistical Mechanics, Lecture Notes in Physics*. Springer, 2003. 17

[63] James Holland Jones and Mark S. Handcock. An assessment of preferential attachment as a mechanism for human sexual network formation, 2002. 17

[64] Marta Mulawa, Thespina J. Yamanis, Lauren M. Hill, Peter Balvanz, Lusajo J. Kajula, and Suzanne Maman. Evidence of social network influence on multiple {HIV} risk behaviors and normative beliefs among young tanzanian men. *Social Science Medicine*, 153:35 – 43, 2016. 17

[65] Francesco Calderoni, Domenico Brunetto, and Carlo Piccardi. Commu-

nities in criminal networks: A case study. *Social Networks*, 48:116 – 125, 2017. 17

[66] Charles Z. Marshak, M. Puck Rombach, Andrea L. Bertozzi, and Maria R. D'Orsogna. Growth and containment of a hierarchical criminal network. *Phys. Rev. E*, 93:022308, Feb 2016. 17

[67] Z. Su, Q. Xu, H. Zhu, and Y. Wang. A novel design for content delivery over software defined mobile social networks. *IEEE Network*, 29(4):62–67, July 2015. 17

[68] Z. Su, Q. Xu, and Q. Qi. Big data in mobile social networks: a qoe-oriented framework. *IEEE Network*, 30(1):52–57, January 2016. 17

[69] Huan Zhou, Linping Tong, Shouzhi Xu, Chungming Huang, and Jialu Fan. Predicting temporal centrality in opportunistic mobile social networks based on social behavior of people. *Personal and Ubiquitous Computing*, 20(6):885–897, 2016. 17

[70] M. E. J. Newman. Random graphs as models of networks. *eprint arXiv:cond-mat/0202208*, February 2002. 17, 21

[71] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97, Jan 2002. 26

[72] Petter Holme and Fredrik Liljeros. Birth and death of links control disease spreading in empirical contact networks. 4:4999+, May 2014. 28

[73] H.R. de Sa and R.B.C. Prudencio. Supervised link prediction in weighted networks. In *Neural Networks (IJCNN), The 2011 International Joint Conference on*, pages 2281–2288, 31 2011-Aug. 5. 30

[74] T. Sørensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Biol. Skr.*, 5:1–34, 1948. 32

[75] Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18:39–43, 1953. 32

[76] M. E. J. Newman. Clustering and preferential attachment in growing networks. *Phys. Rev. E*, 64:025102, Jul 2001. 33

[77] Fei Gao, Katarzyna Musial, Colin Cooper, and Sophia Tsoka. Link prediction methods and their accuracy for different social networks and network metrics. *Scientific Programming*, 2015, 2015. 33, 173

[78] Q. Ou, Y. D. Jin, T. Zhou, B. H. Wang, and B. Q. Yin. Power-law strength-degree correlation from resource-allocation dynamics on weighted networks. *Phys. Rev. E*, 75:021102, 2007. 34

[79] Peng Wang, Baowen Xu, Yurong Wu, and Xiaoyu Zhou. Link prediction in social networks: the state-of-the-art. *CoRR*, abs/1411.5118, 2014. 34

[80] E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai, and A.L. Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551, 2002. 34

[81] E. A. Leicht, Petter Holme, and M. E. J. Newman. Vertex similarity in networks. *Phys. Rev. E*, 73:026120, Feb 2006. 35

[82] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pages 11–15, Pasadena, CA USA, August 2008. 35

[83] G. Bianconi and A. L. Barabsi. Competition and multiscaling in evolving networks. *Europhysics Letters*, 54:436–442, 2001. 38

[84] M. Budka, K. Musial, and K. Juszczyszyn. Predicting the evolution of social networks: Optimal time window size for increased accuracy. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 21–30, Sept 2012. 51, 63, 88

[85] Brani Vidakovic. *Statistics for Bioengineering Sciences*. Springer New York, 2011. 54

[86] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, April 1982. 55

[87] Aaron Clauset, Cristopher Moore, and M. E. J. Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453:98–101, 2008. 55

[88] Jrme Kunegis. Konect: the koblenz network collection. In *WWW (Companion Volume)*, pages 1343–1350. International World Wide Web Conferences Steering Committee / ACM, 2013. 56, 67

[89] Przemyslaw Kazienko, Katarzyna Musial, and Aleksander Zgrzywa. Evaluation of node position based on email communication. *Control and Cybernetics*, 38(1):67–86, 2009. 56

[90] Bryan Klimt and Yiming Yang. The Enron corpus: A new dataset for email classification research. In *Proc. European Conf. on Machine Learning*, pages 217–226, 2004. 57

[91] Tore Opsahl and Pietro Panzarasa. Triadic closure in two-mode networks: Redefining the global and local clustering coefficients. *Social Networks*, 34, 2011. 57

[92] Bimal Viswanath, Alan Mislove, Meeyoung Cha, and Krishna P. Gummadi. On the evolution of user interaction in Facebook. In *Proc. Workshop on Online Social Networks*, pages 37–42, 2009. 57, 62

[93] Alan Mislove, Hema Swetha Koppula, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. Growth of the Flickr social network. In *Proc. Workshop on Online Social Networks*, pages 25–30, 2008. 57, 62

[94] Alan Mislove. *Online Social Networks: Measurement, Analysis, and Applications to Distributed Information Systems*. PhD thesis, Rice University, 2009. 57

[95] Twitter network dataset – KONECT, October 2016. 57, 62

[96] Munmun De Choudhury, Yu-Ru Lin, Hari Sundaram, K. Seluk Candan, Lexing Xie, and Aisling Kelliher. How does the data sampling strategy impact the discovery of information diffusion in social media? In *ICWSM*, pages 34–41, 2010. 57

[97] Giulia Berlusconi, Francesco Calderoni, Nicola Parolini, Marco Verani, and Carlo Piccardi. Link prediction in criminal networks: A tool for criminal intelligence analysis. *PLOS ONE*, 11(4):1–21, 04 2016. 66

[98] Joseph L. Rodgers and Alan W. Nicewander. Thirteen Ways to Look at the Correlation Coefficient. *The American Statistician*, 42(1):59–66, 1988. 67

[99] Jrme Kunegis and Julia Preusse. Fairness on the web: Alternatives to the power law. In *Proc. Web Science Conf.*, 2012. 69

[100] Mark E. J. Newman, Albert L. Barabási, and Duncan J. Watts, editors. *The Structure and Dynamics of Networks*. Princeton University Press, 2006. 71

[101] Reza Bakhshandeh, Mehdi Samadi, Zohreh Azimifar, and Jonathan Schaeffer. Degrees of separation in social networks. In *SOCS*, 2011. 71

[102] Fei Gao, Katarzyna Musial, Colin Cooper, and Sophia Tsoka. Link prediction methods and their accuracy for different social networks and network metrics. *Scientific Programming*, 2015:172879:1–172879:13, 2015. 83

[103] Michael Grant and Stephen Boyd. Graph implementations for nonsmooth convex programs. In V. Blondel, S. Boyd, and H. Kimura, editors, *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pages 95–110. Springer-Verlag Limited, 2008. 86

[104] Michael Grant and Stephen Boyd. Cvx: Matlab software for disciplined convex programming, version 2.1. `http://cvxr.com/cvx`, mar 2014. 86

[105] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: An open source software for exploring and manipulating networks, 2009. 147

[106] V.D. Blondel, J.L. Guillaume, R. Lambiotte, and E.L.J.S. Mech. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, page P10008, 2008. 147, 166

[107] Zuxi Wang, Yao Wu, Qingguang Li, Fengdong Jin, and Wei Xiong. Link prediction based on hyperbolic mapping with community structure for

complex networks. *Physica A: Statistical Mechanics and its Applications*, 450:609 – 623, 2016. 160

[108] Di Jin, Mengdi Wang, and Yu-Ru Lin. *TeleLink: Link Prediction in Social Network Based on Multiplex Cohesive Structures*, pages 174–185. Springer International Publishing, Cham, 2016. 160

[109] Jingyi Ding, Licheng Jiao, Jianshe Wu, Yunting Hou, and Yutao Qi. Prediction of missing links based on multi-resolution community division. *Physica A: Statistical Mechanics and its Applications*, 417:76 – 85, 2015. 160

[110] Sucheta Soundarajan and John Hopcroft. Using community information to improve the precision of link prediction methods. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12 Companion, pages 607–608, New York, NY, USA, 2012. ACM. 160, 161, 165

[111] Katarzyna Musiał and Krzysztof Juszczyszyn. *Properties of Bridge Nodes in Social Networks*, pages 357–364. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. 162

[112] Santo Fortunato and Darko Hric. Community detection in networks: A user guide. *CoRR*, abs/1608.00163, 2016. 166

[113] F. Gao and K. Musial-Gabrys. Hybrid structure-based link prediction model. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 1221–1228, Aug 2016. 174